



Accelerating Insight Creation with Data Mesh

August 2024

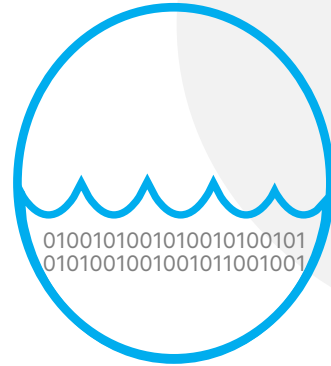
Journey to a Democratized Mesh: Overview



Leading Financial
Research and
Data Solutions



1984



Centralized Data
Lake
and Lakehouse



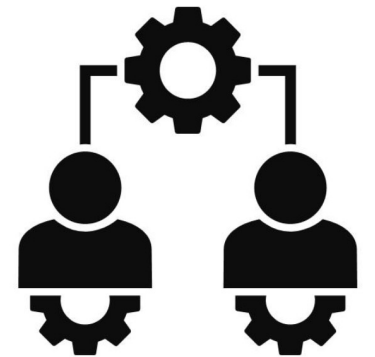
2018



Self-Service
Platform



2021



Data Mesh
Operating Model



2024

About Morningstar

Empowering Investor Success

Asset Managers



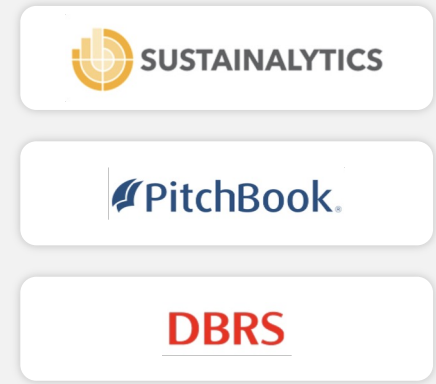
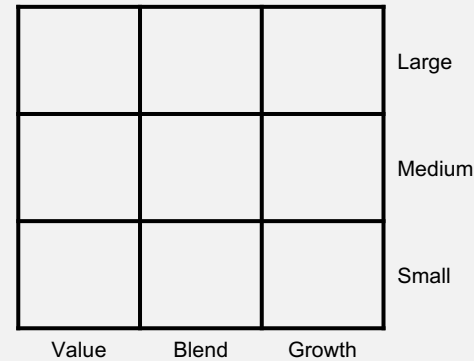
Investors

Advisors

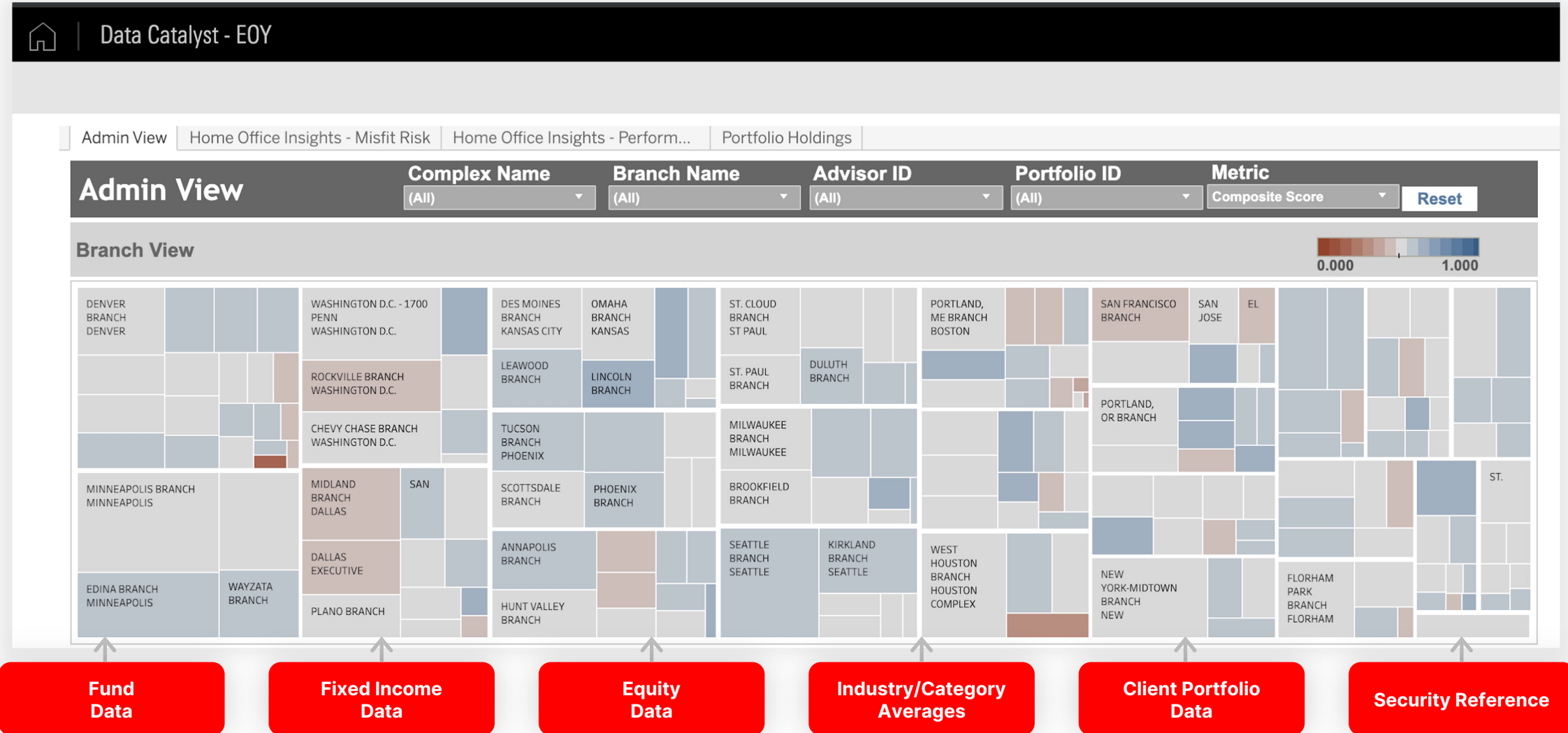
40 Years and Counting



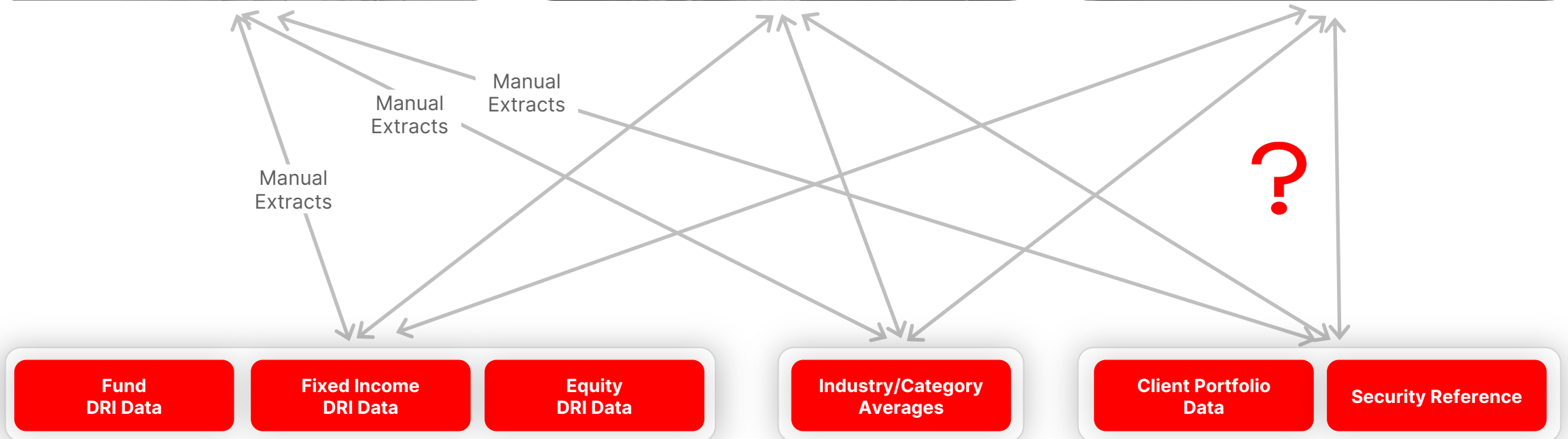
Morningstar Style Box



Integrating Diverse Datasets: Morningstar Data Catalyst

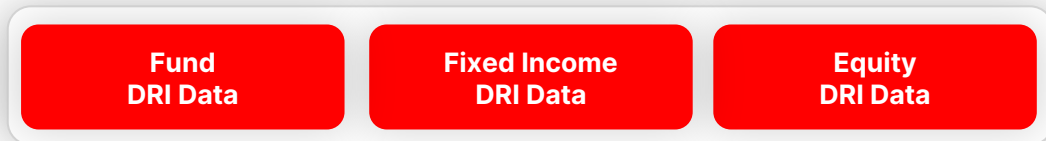
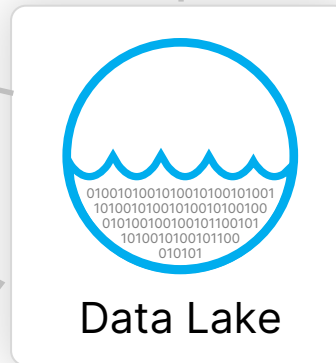


Fragmented Data Sources -> Handoffs



Started our data journey by centralizing data

Morningstar's data lake



Making Data Consumer Friendly

Morningstar's Lakehouse

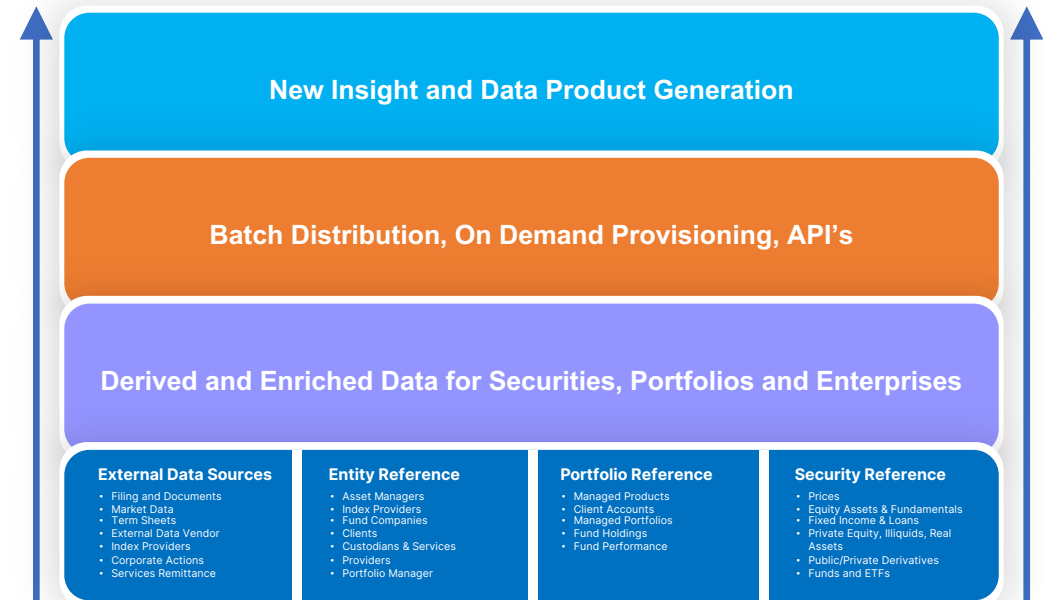
Redshift

addresses query speed



Modeling

addresses data ambiguity

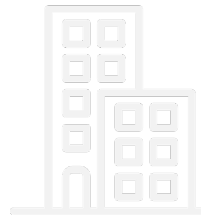


Realization of Problem - Technology Alone is not the Solution



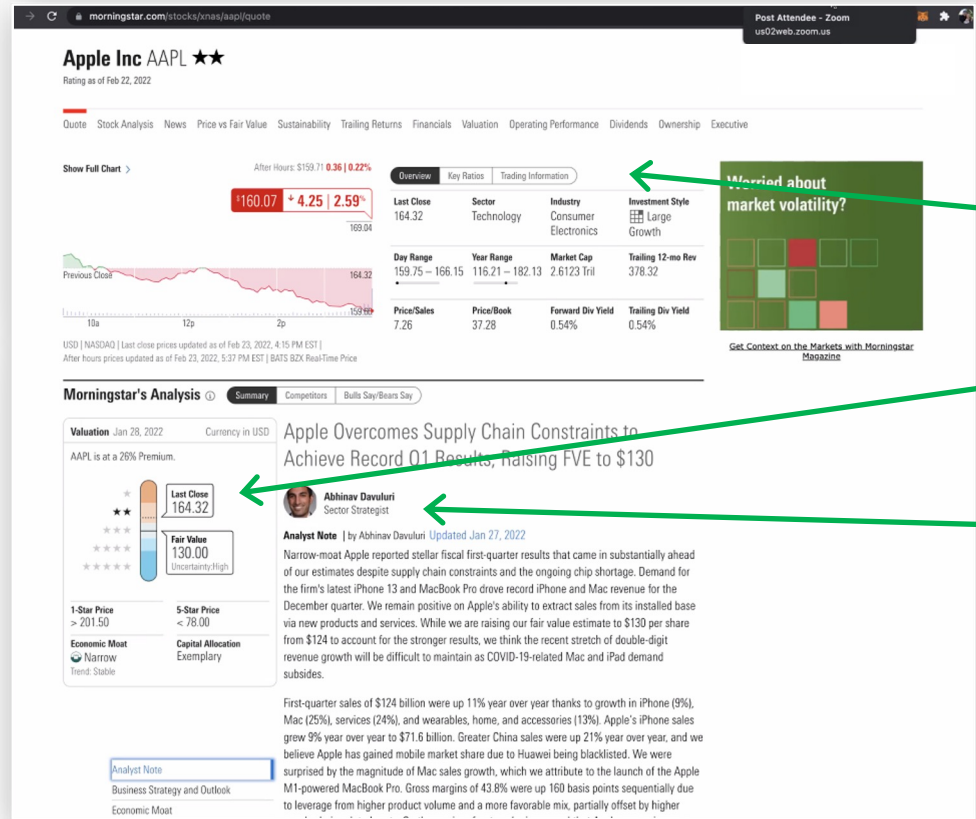
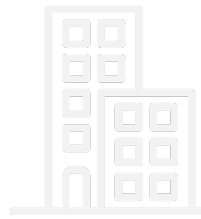
Technology

Supported by:
People and Research



People & Research

Supported by:
Technology




Technology

Data

Research

People

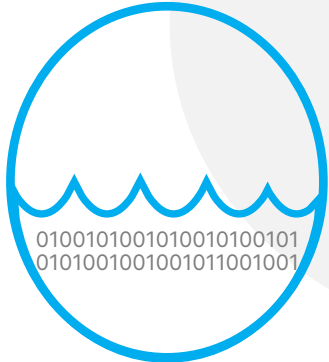
Journey to a Democratized Mesh: Self-Service



Leading Financial Research and Data Solutions

→


1984



Centralized Data Lake and Lakehouse

→

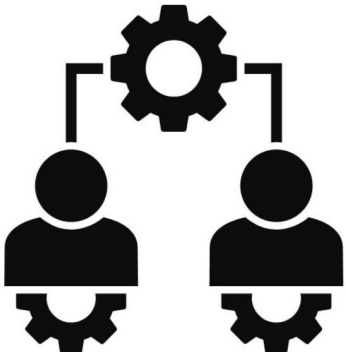
2018



Self-Service Platform

→

2021



Data Mesh Operating Model

→

2024

Problem - How Do We Scale Through Our People?

	Data Ingestion	Data Quality	Data Modeling
People Processes ▶	<u>Content specialists and developers</u> create ETL pipelines..	<u>Business analyst, domain SME's and developers</u> set validation rules...	<u>Content specialists and DBAs</u> define modeling...
How? ▶	... back and forth communication and handoffs		
Scaling Issues ▶	1000s of lines of ETL code to write	Validation rule volume is unruly to manage <ul style="list-style-type: none">• Sheer volume is an issue• But validation rules can change as market does• 10's of 1000's of tests	Keeping methodology and code in sync is problematic 1000+ views

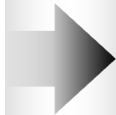
Benefits - Data Ingestion

Self-service beyond the specialists



```
AWS Glue > Developer Guide
root
|-- role: string
|-- seats: int
|-- org_name: string
|-- links: array
| | |-- element: struct
| | | | |-- note: string
| | | | |-- url: string
| | |-- type: string
| |-- sort_name: string
| |-- area_id: string
| |-- images: array
| | |-- element: struct
| | | | |-- url: string
| | |-- on_behalf_of_id: string
| | |-- other_names: array
| | | | |-- element: struct
| | | | | | |-- note: string
| | | | | | |-- name: string
| | | | | | |-- lang: string
| | |-- contact_details: array
| | | | |-- element: struct
| | | | | | |-- type: string
| | | | | | |-- value: string
| |-- name: string
| |-- birth_date: string
| |-- organization_id: string
| |-- gender: string
| |-- classification: string
| |-- death_date: string
| |-- legislative_period_id: string
| |-- identifiers: array
| | |-- element: struct
```

**AWS Glue
Code + Airflow
Complexity**



The screenshot shows the etleap 'Create Pipeline' interface. At the top, there's a navigation bar with 'Home', 'Activities', 'Connections', and 'Search'. Below that, a 'Create Pipeline' header has a 'BACK' button and a 'CREATE' button. The main area is divided into a table and a script editor.

	event_type	referrer	ipaddress	joindate	cookie	event
1	Click	https://www.doggyswag.com/about	67.234.12.43	2016-03-19	NL9WBHYZLCL9ZX1UC6V8	2019-05-07T18:42:55-07:00
2	Click	https://www.doggyswag.com/catalog/rubberballs	67.234.12.43	2019-03-16	GTJWE050Q3QPSFXYDX	2019-05-07T16:24:30-07:00
3	PageLoad	null	67.234.12.43	2015-12-12	90JOCUEN9XF8HECD37FA	2019-05-07T05:52:42-07:00
4	PageLoad	null	213.34.56.96	2017-01-13	M0TG7RXXN80C684C56F	2019-05-07T16:43:31-07:00
5	PageLoad	null	39.43.233.2	2016-03-10	SKRMQHK6T1JFDHDCWE9V	2019-05-07T07:31:58-07:00
6	Click	https://www.doggyswag.com/catalog/plush-toys	213.34.56.96	2015-04-30	ZKXFPKZYKPMHP5V8KGI	2019-05-07T20:35:39-07:00
7	PageLoad	null	67.234.12.43	2015-12-09	CE7MXXONS5LXZ2N2MB98	2019-05-07T17:37:53-07:00
8	PageLoad	null	54.193.34.3	2015-02-19	ELD3DCULTHFFKL7BMAJW	2019-05-07T00:41:20-07:00
9	Click	https://www.doggyswag.com/account	12.13.200.32	2016-01-08	HMP2QSGGAZXXHR2KNTDO	2019-05-07T12:03:39-07:00
10	PageLoad	null	54.193.34.3	2016-01-05	WGAWS0WQZKNOF39ZAME	2019-05-07T07:08:47-07:00
11	PageLoad	null	39.43.233.2	2017-12-29	6R3KQJAF60YB8CFQJRX	2019-05-07T19:23:23-07:00
12	PageLoad	null	39.43.233.2	2018-12-02	LHLIR80IAP0FXJJYH3UH	2019-05-07T17:49:57-07:00

The script editor on the right contains the following steps:

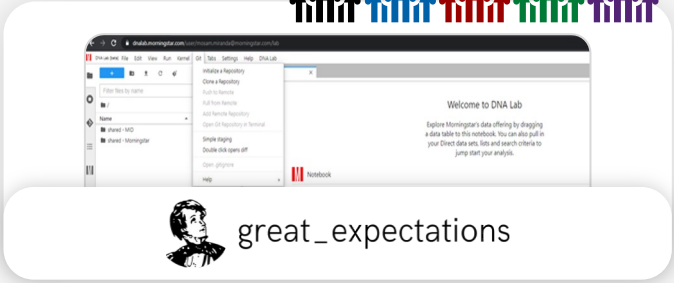
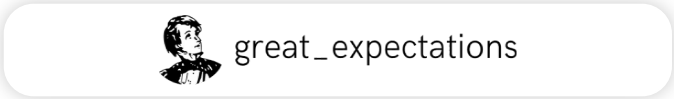
- 1 Split [data] repeatedly on newLine
- 2 Split [data] once on space character
- 3 Rename column split to event_type
- 4 Flatten JSON object in split1 with all fields
- 5 Interpret joindate as date with format MM/dd/yyyy

Buttons for '+ ADD SCRIPT STEP' and 'NEXT' are visible at the bottom of the script editor.



Benefits - Data Quality

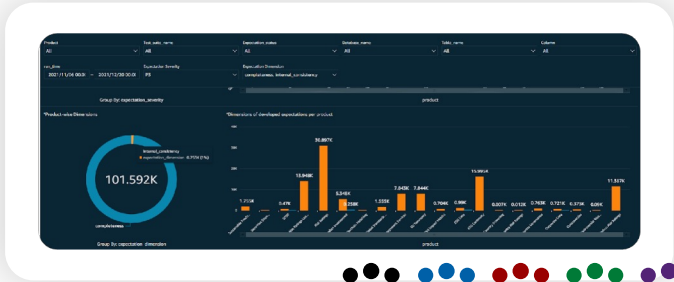
Increased input, visibility, integration



JupyterLab

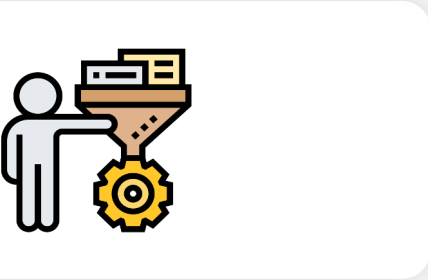


Reporting Dashboard

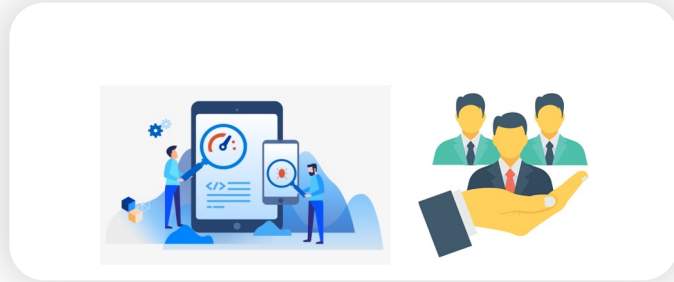


Update Test Suite

OR



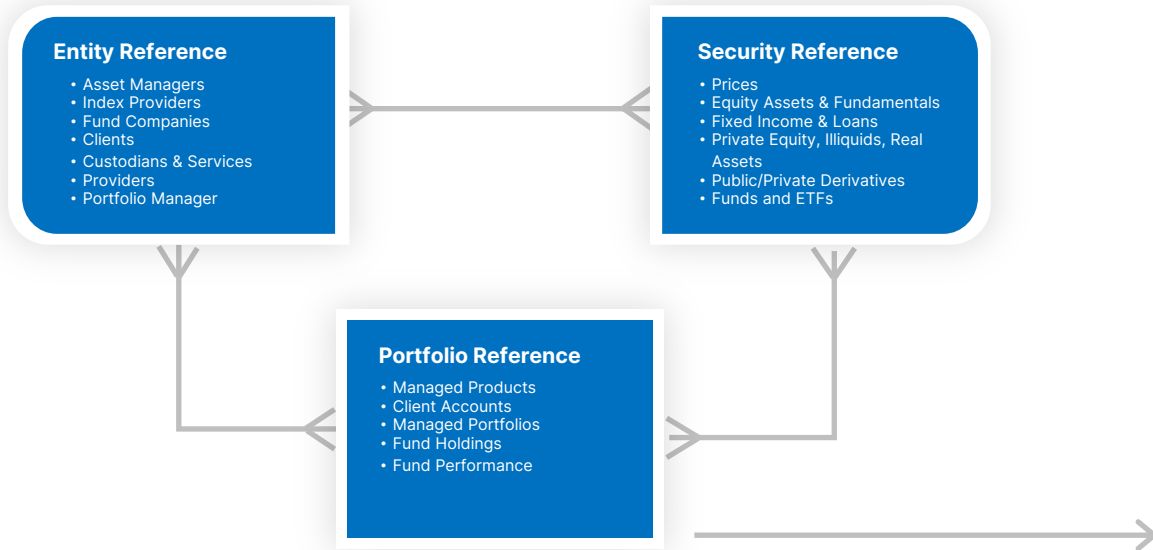
Fix



Data SME / Client

Benefits - Data Modeling

Data SMEs informing data relationships



Tie model development with methodology documentation

2) `pai_dictionary` implemented as a reference guide
use it to create a dictionary that will include the lakehouse target table and target column

```

[10]: pai_dictionary = pd.read_csv('pai_dictionary.csv', encoding='unicode_escape')
      pai_dictionary
      indicator_id = {}
      for i in pai_dictionary['indicator_id']:
          if pd.isnull(i) is True:
              indicator_id.append(i)
          else:
              indicator_id.append(int(i))
      pai_dictionary['indicator_id'] = indicator_id
  
```

	Lakehouse Consumer View Target Table	Lakehouse Consumer View Target Column	Data Type	Primary Key	Is	definition	Fieldtype	Field Value Range	Data Lak
0	principal_adverse_indicators_mandatory_for_ent...	entry_id	char(10)	True	True	Lakehouse Unique Entity Identifier	Character	NaN	company_esp_prd.sustainable_fi...
1	principal_adverse_indicators_mandatory_for_ent...	date_valid_from	date	True	NaN	Date	Date	NaN	company_esp_prd.sustainable_fi...
2	principal_adverse_indicators_mandatory_for_ent...	scope_2_greenhouse_gas_emissions	decimal(17,2)	NaN	NaN	Carbon - Scope 1 Emissions	Numeric	NULLmin. 0.00 1pmax. 999999999.99	company_esp_prd.sustainable_fi...
3	principal_adverse_indicators_mandatory_for_ent...	scope_2_greenhouse_gas_emissions	decimal(17,2)	NaN	NaN	Carbon - Scope 2 Emissions	Numeric	NULLmin. 0.00 1pmax. 999999999.99	company_esp_prd.sustainable_fi...
4	principal_adverse_indicators_mandatory_for_ent...	scope_3_greenhouse_gas_emissions	decimal(17,2)	NaN	NaN	Carbon - Scope 3 Emissions	Numeric	NULLmin. 0.00 1pmax. 999999999.99	company_esp_prd.sustainable_fi...
...
62	principal_adverse_indicators_additional_social...	human_rights_score	decimal(9,4)	NaN	NaN	This metric assesses the average human rights ...	Numeric	NULLmin. 0.0000 max. 100.0000	country_esp_prd.sustainable_fi...
63	principal_adverse_indicators_additional_social...	control_of_corruption	decimal(9,4)	NaN	NaN	IC - Control of Corruption Score	Numeric	NULLmin. 0.0000 max. 100.0000	country_esp_prd.sustainable_fi...
64	principal_adverse_indicators_additional_social...	non-cooperative_tax_jurisdictions	boolean	NaN	NaN	Binary flag signaling jurisdictions included	Binary	NULLTRUEFALSE	country_esp_prd.sustainable_fi...
65	principal_adverse_indicators_additional_social...	political_stability	decimal(9,4)	NaN	NaN	IC - Political Stability Score	Numeric	NULLmin. 0.0000 max. 100.0000	country_esp_prd.sustainable_fi...
66	principal_adverse_indicators_additional_social...	rule_of_law	decimal(9,4)	NaN	NaN	IC - Rule of Law Score	Numeric	NULLmin. 0.0000 max. 100.0000	country_esp_prd.sustainable_fi...

67 rows x 11 columns

3) Apply logic to create target tables from source tables

3-1) create de-duplicated list of entry_id and date_valid from

```

[12]: company_profile_dateval=company_sustainable_finance_disclosure_regulation_profiles_indicators[['entry_id', 'date_valid_from']].drop_duplicates()
  
```

3-2) Take data lake source tables and convert vertical data into horizontal data, store data into dictionary organized by target tables and their respective target columns. Referencing the `pai_dictionary` to perform the task.

```

[14]: dictionary = {}
      for i in pai_dictionary['Lakehouse Consumer View Target Table'].unique():
          sub_dictionary = {}
          for j, k, l in zip(pai_dictionary.loc[pai_dictionary['Lakehouse Consumer View Target Table'] == i]['indicator_id'], pai_dictionary.loc[pai_dictionary['Lakehouse Consumer View Target Table'] == i]['entry_id'], pai_dictionary.loc[pai_dictionary['Lakehouse Consumer View Target Table'] == i]['date_valid_from']):
              sub_dictionary[j] = {'entry_id': k, 'date_valid_from': l}
          dictionary[i] = sub_dictionary
  
```

Making Data Consumer Friendly

Morningstar's Lakehouse

Redshift

addresses query speed



Data Lake
etleap

Modeling

addresses data ambiguity

Entity Reference

- Asset Managers
- Index Providers
- Fund Companies
- Clients
- Custodians & Services
- Providers
- Portfolio Manager

Portfolio Reference

- Managed Products
- Client Accounts
- Managed Portfolios
- Fund Holdings
- Fund Performance

Security Reference

- Prices
- Equity Assets & Fundamentals
- Fixed Income & Loans
- Private Equity, Illiquids, Real Assets
- Public/Private Derivatives
- Funds and ETFs

Making Data Consumer Friendly

Morningstar's Lakehouse

Redshift

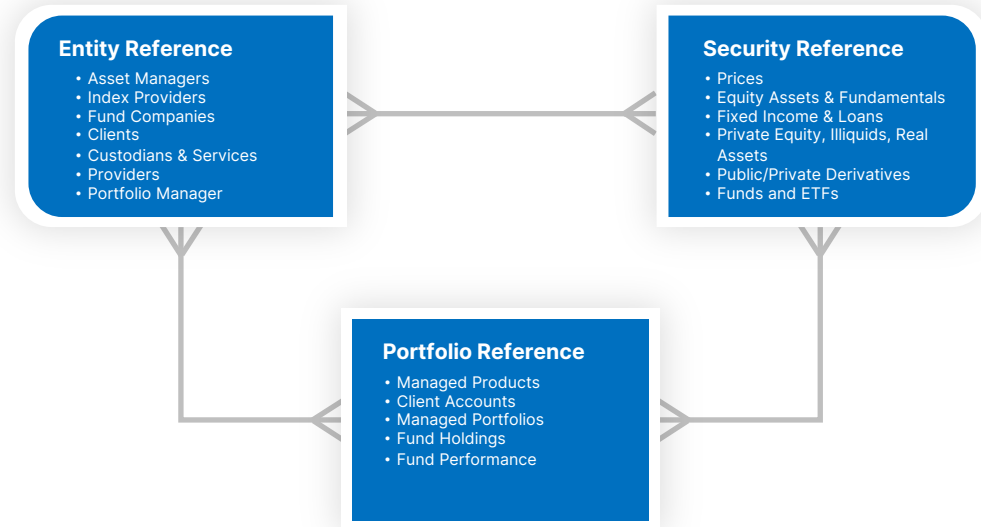
addresses query speed



Data Lake
etleap

Modeling

addresses data ambiguity

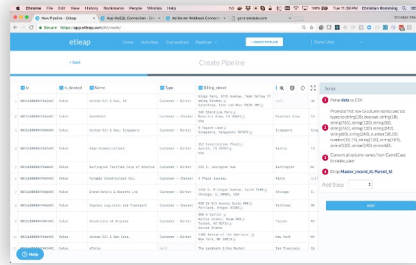


Benefits - Scaling Through Self-Service

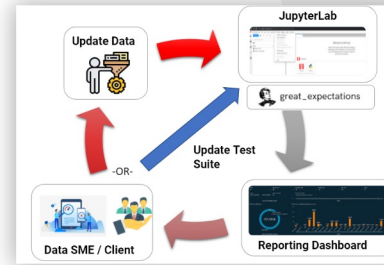
Solution



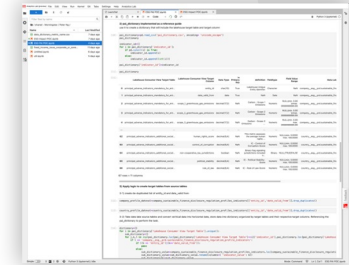
Data Ingestion



Data Quality



Data Modeling



Results



ETL onboarding from weeks to minutes

Total production pipelines: **2900 and counting**


Dozens of teams producing ingestion

Current tests: **24k+**
Current executions: **15mil**, with **2.2%** failure rate

Models: **600+ by core team**, POC'ing Finance team

BA's: **Hundreds** scattered across the organization

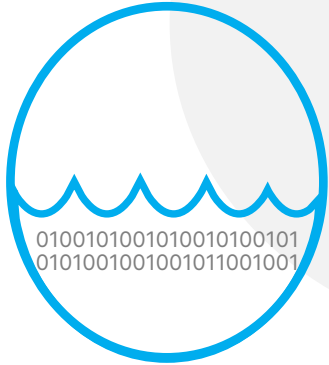
Journey to a Democratized Mesh: Self-Service



Leading Financial Research and Data Solutions

→


1984



Centralized Data Lake and Lakehouse

→

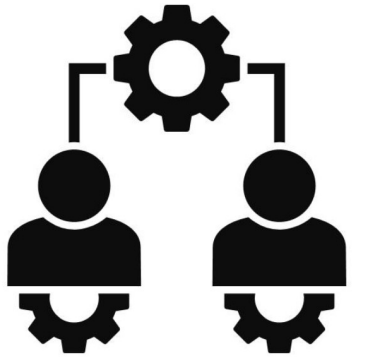
2018



Self-Service Platform

→

2021



Data Mesh Operating Model

→

2024

Hueler + Morningstar Analytics

Morningstar Acquires Hueler Analytics' Stable Value Data and Index

The addition of stable-value product data to Morningstar data increases visibility for assets in defined-contribution (DC) retirement plans



NEWS PROVIDED BY

[Morningstar, Inc.](#) →

03 Feb, 2020, 09:00 ET

SHARE THIS ARTICLE

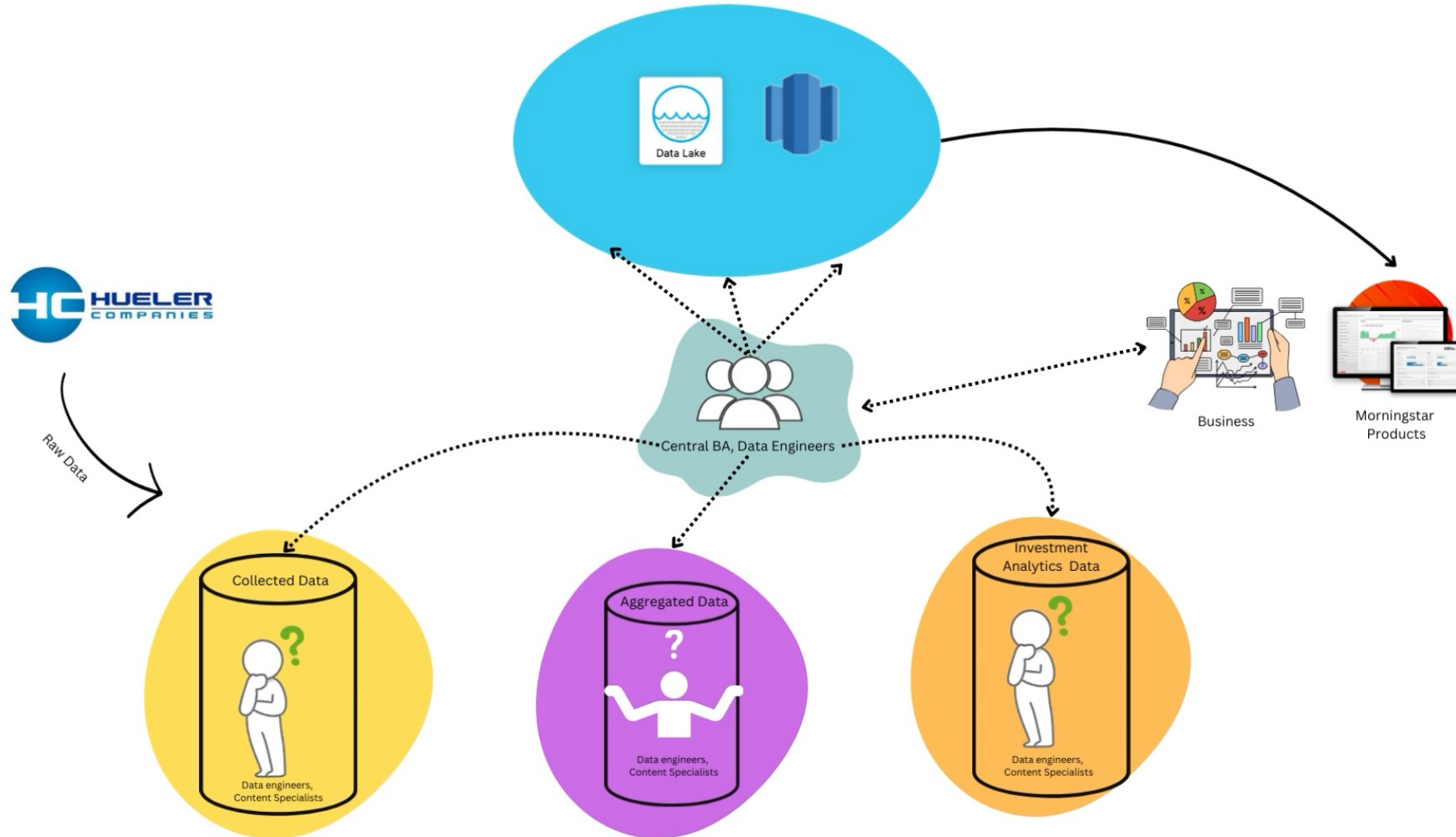


CHICAGO, Feb. 3, 2020 /PRNewswire/ -- [Morningstar, Inc.](#) (Nasdaq: [MORN](#)), a leading provider of independent investment research, today announced its acquisition of [Hueler Analytics'](#) Stable Value Comparative Universe Data and Stable Value Index. Terms of the transaction were not disclosed.

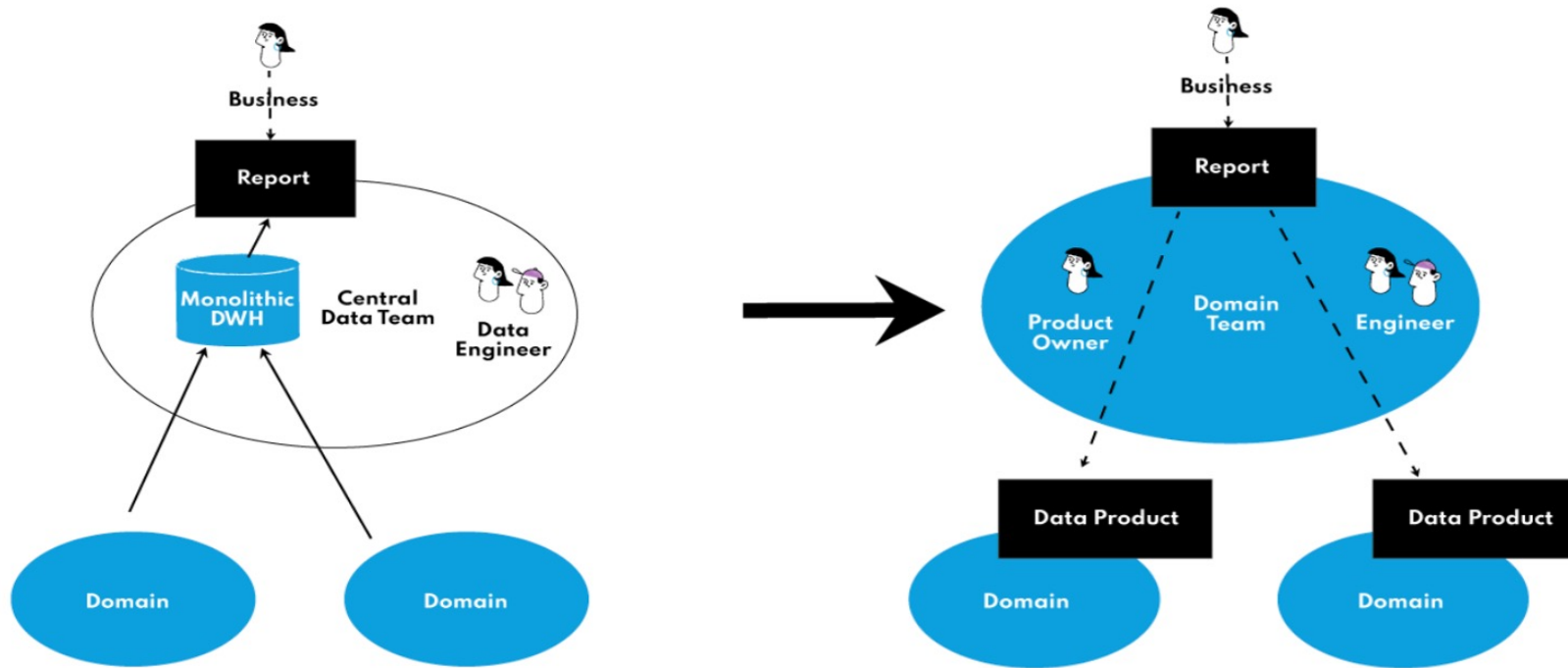
Founded in 1987 in Minneapolis, Minn., Hueler is the premiere independent data and research firm providing reporting and systems designed for the annuity and stable-value marketplace. Hueler Analytics' distribution encompasses advisors, investment managers, product providers, plan fiduciaries, and consultants. Hueler Analytics' Stable Value Comparative Universe Data provides broad market coverage of stable-value investments, including stable-value pooled funds, insurance



A siloed operating model

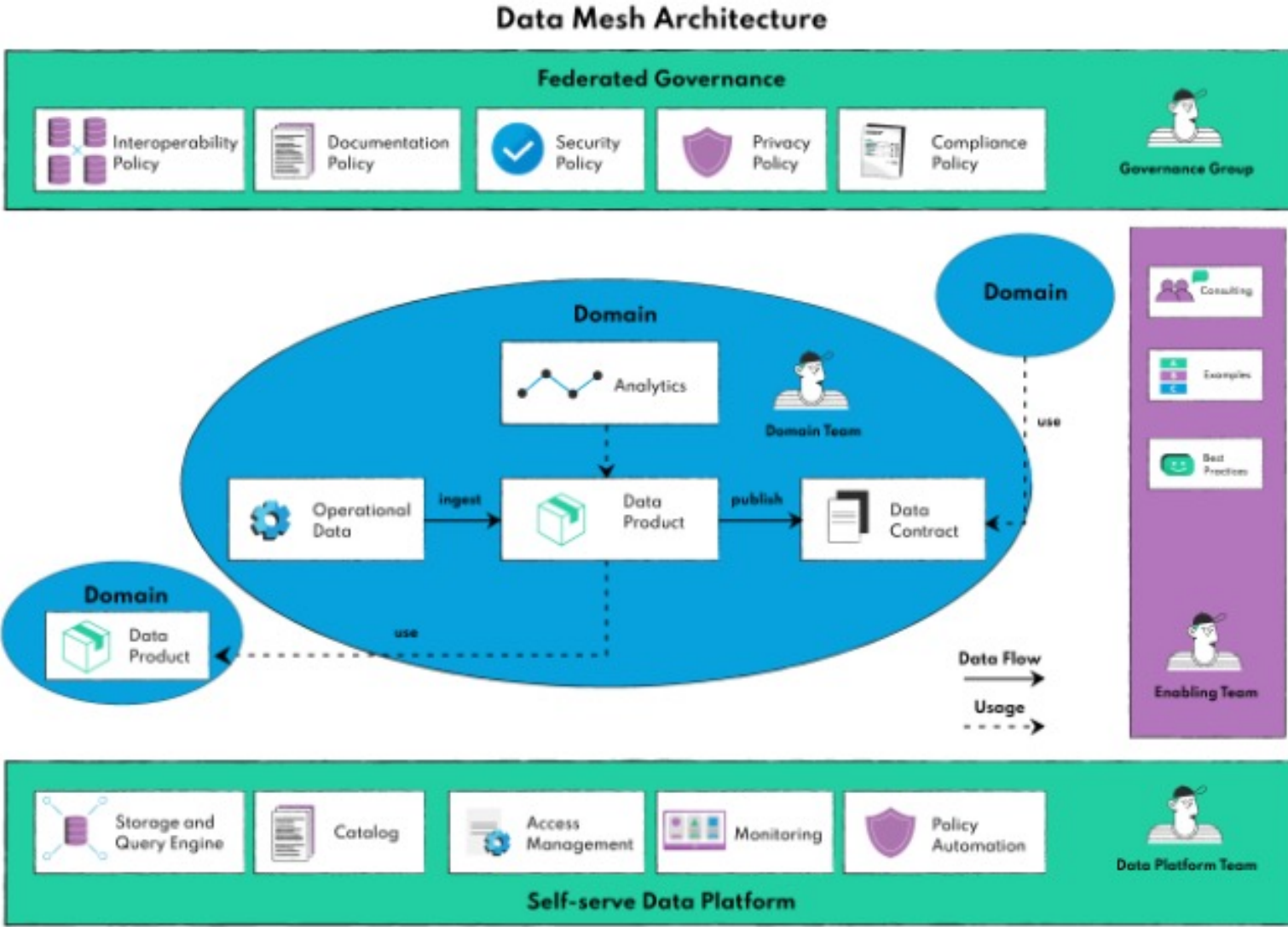


Introduce data mesh

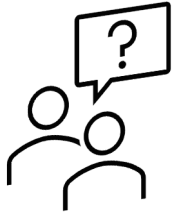


“Data mesh is a decentralized sociotechnical approach to share, access, and manage analytical data in complex and large-scale environments—within or across organizations”
- Dehghani, Zhamak

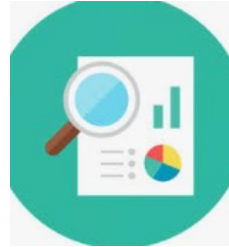
Solution - Data Mesh Architecture



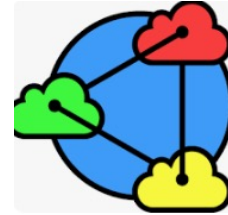
Key Lessons from lack of data mesh



Siloed Data



Lack of transparency in
Data



Lack of consistency in
Data



Lack of true ownership

Our approach

Progress is impossible without change, and those who cannot change their minds cannot change anything
– George Bernard Shaw

Key metrics to measure success



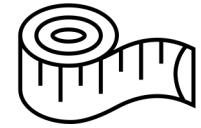
Type of data and how much we absorb



Of the data we have, what is investable

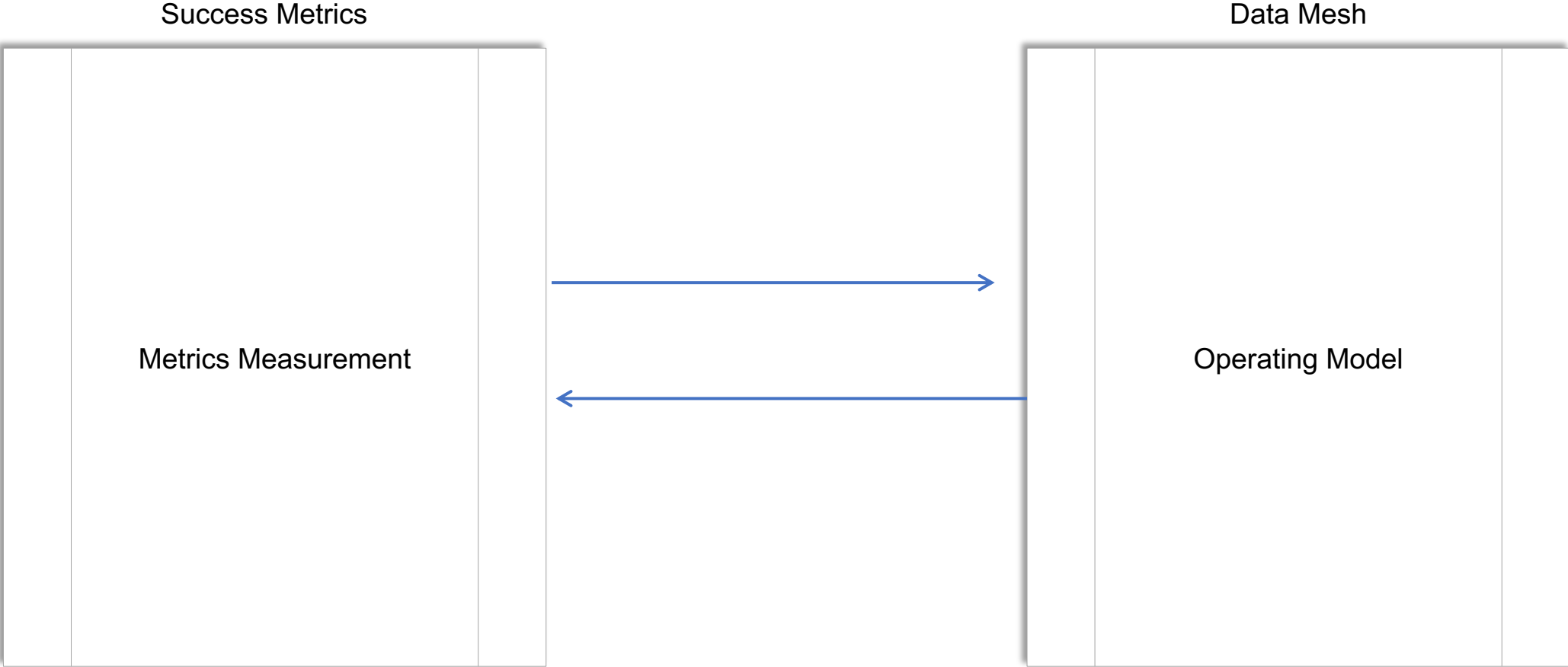


Costs

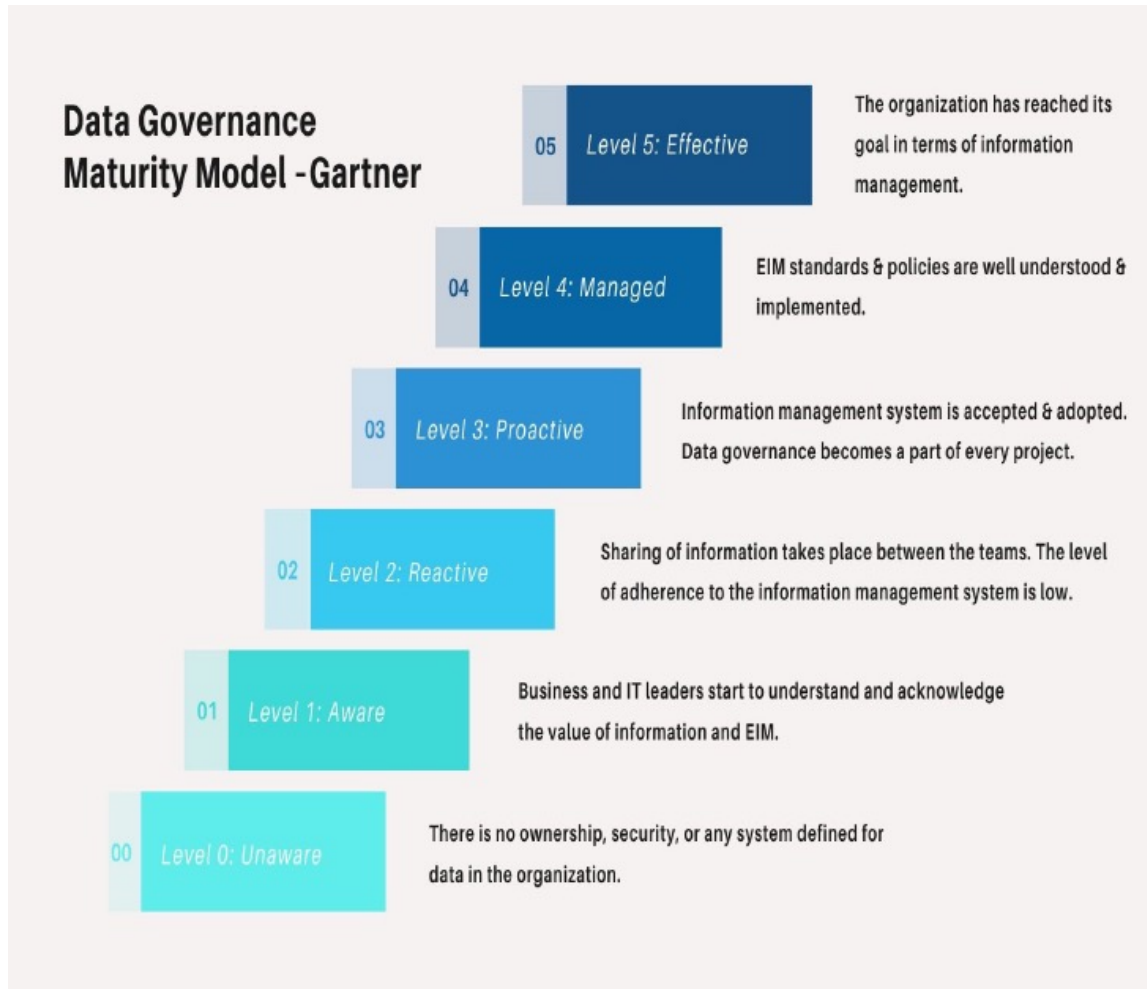


Measuring the Consistency of Data

Partnering Metrics with Mesh



Measuring Maturity



An Example – Data Quality (Availability)

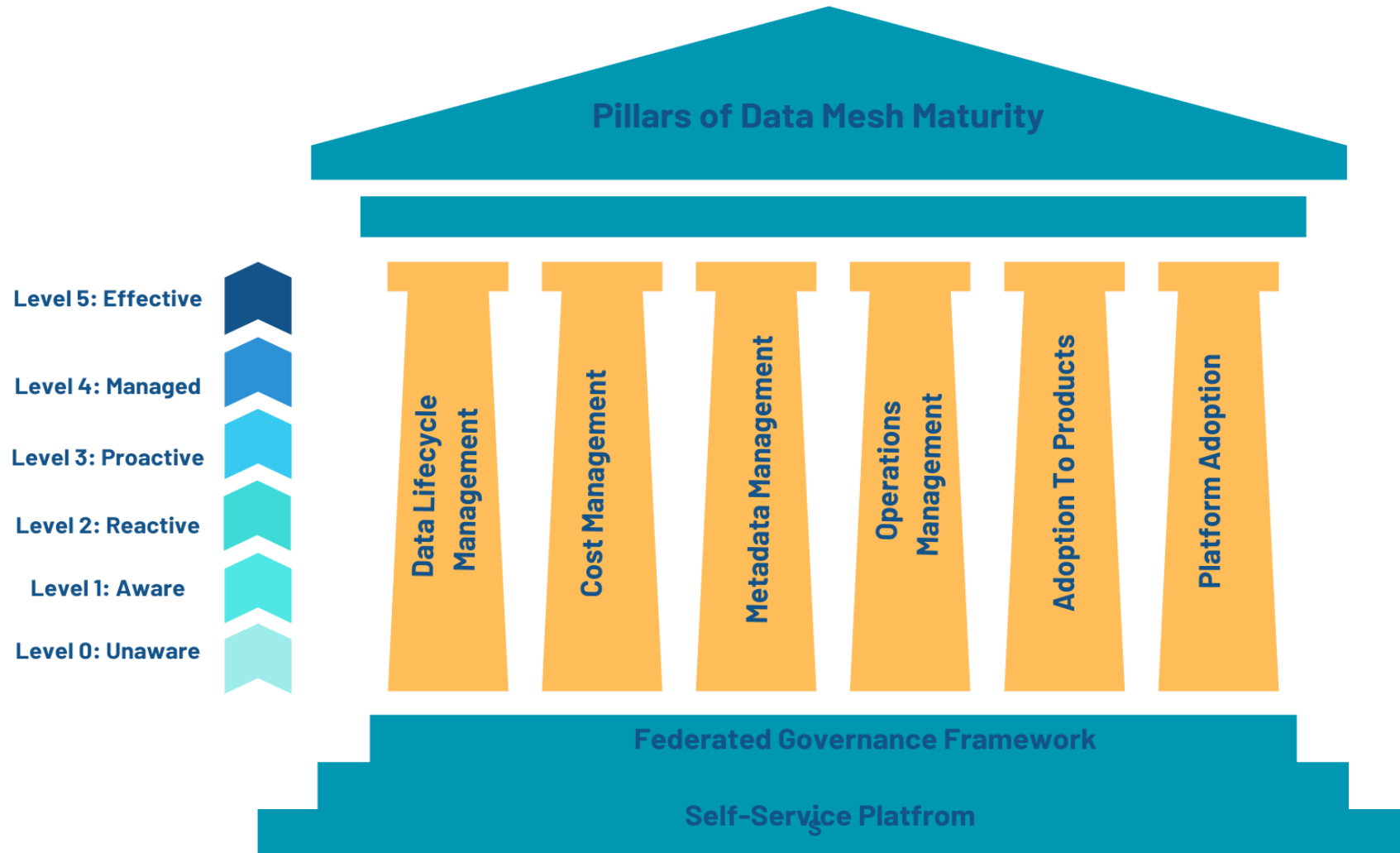
Level n+1: my data is available in all software products

Level n: my data is available in Lakehouse

Level 1: I am verifying the accuracy of my data in the lake

Level 0: my data is being published to data lake

Our Maturity Model



Measuring maturity



END GOAL:
measure
domains/product
maturity across
against
cost/value of
procuring the
data

Changes in responsibilities within domain teams

Traditional Agile Product Owner	Data Product Owner
What are the workflows I need to build for my users?	What are the data views I need to build for my users?
This user wants access to my software product – let me ensure they have an account	This user wants access to my data product – let me ensure they are entitled properly
I will sign-off after every release my workflows are still functioning properly	I will sign-off after every release my data exists in all products properly
I need to open a P1 because no users can login to my platform	I need to open a P1 because my latest data is not available for users in Lakehouse to query

Improvements and Benefits

**Democratizing Data /
Promoting a Culture of
Data Literacy**

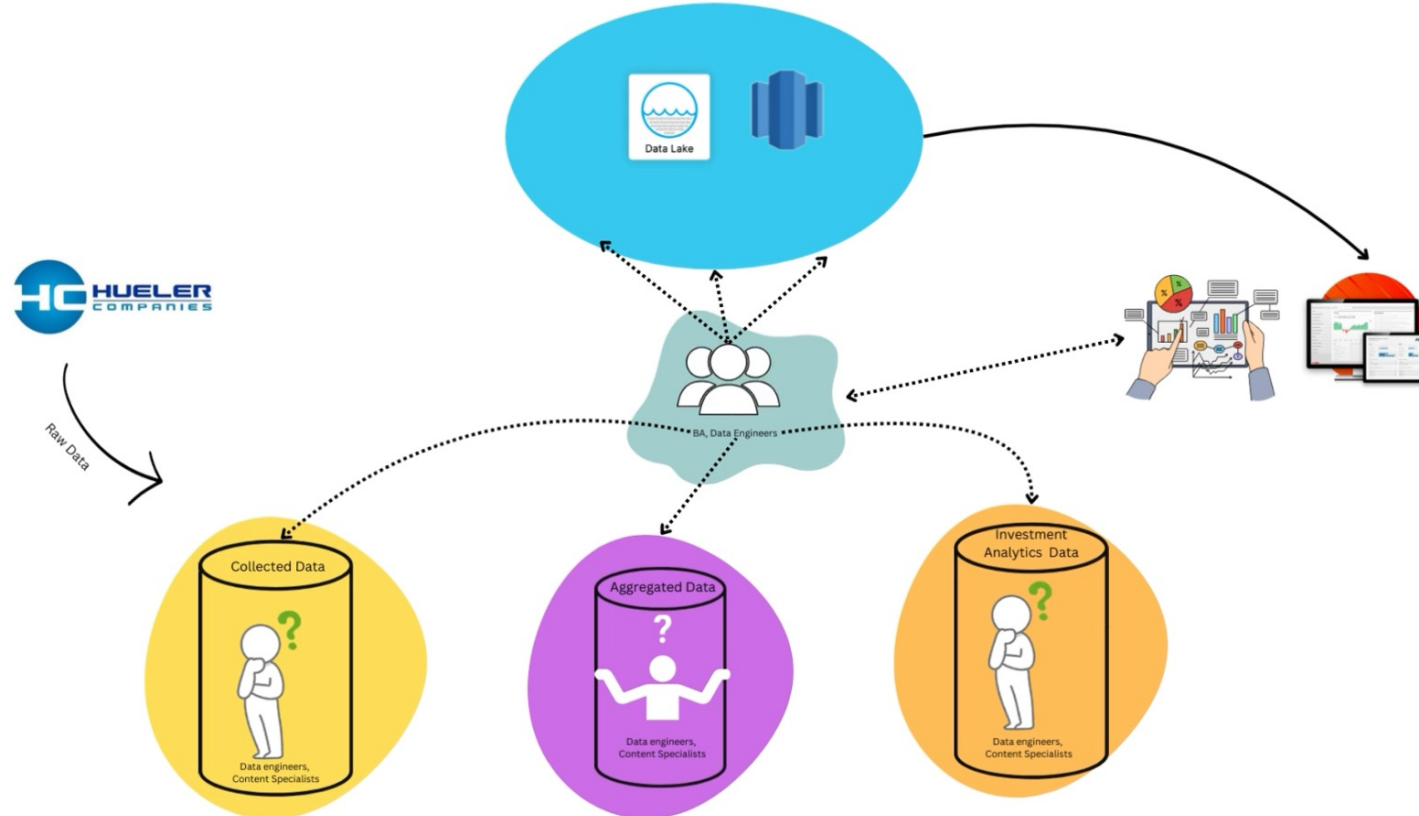
**Scalable, Agile and
Autonomous**

Enhanced Data Quality

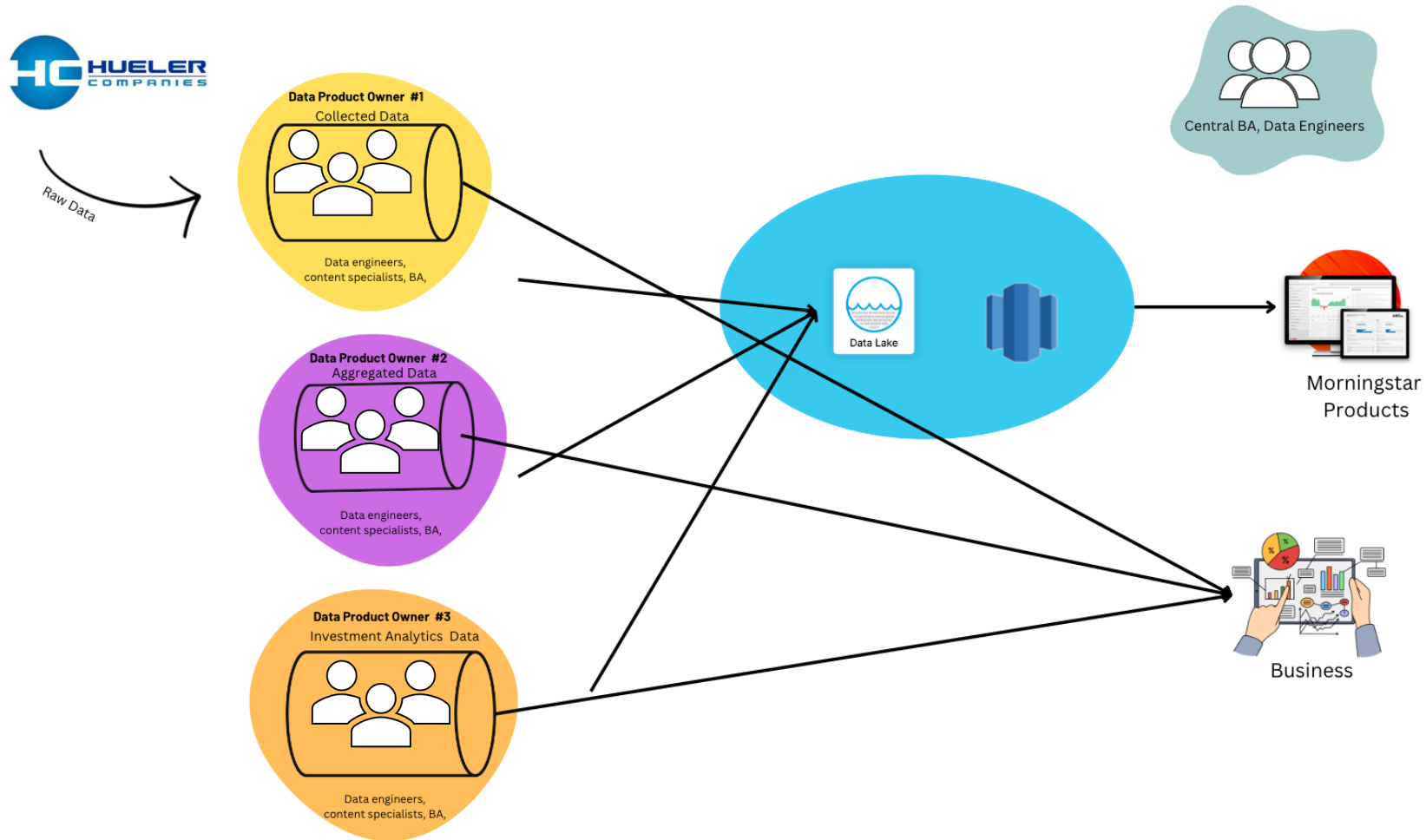
Foster Innovation

**Reduction in repetitive
costs across the
organization**

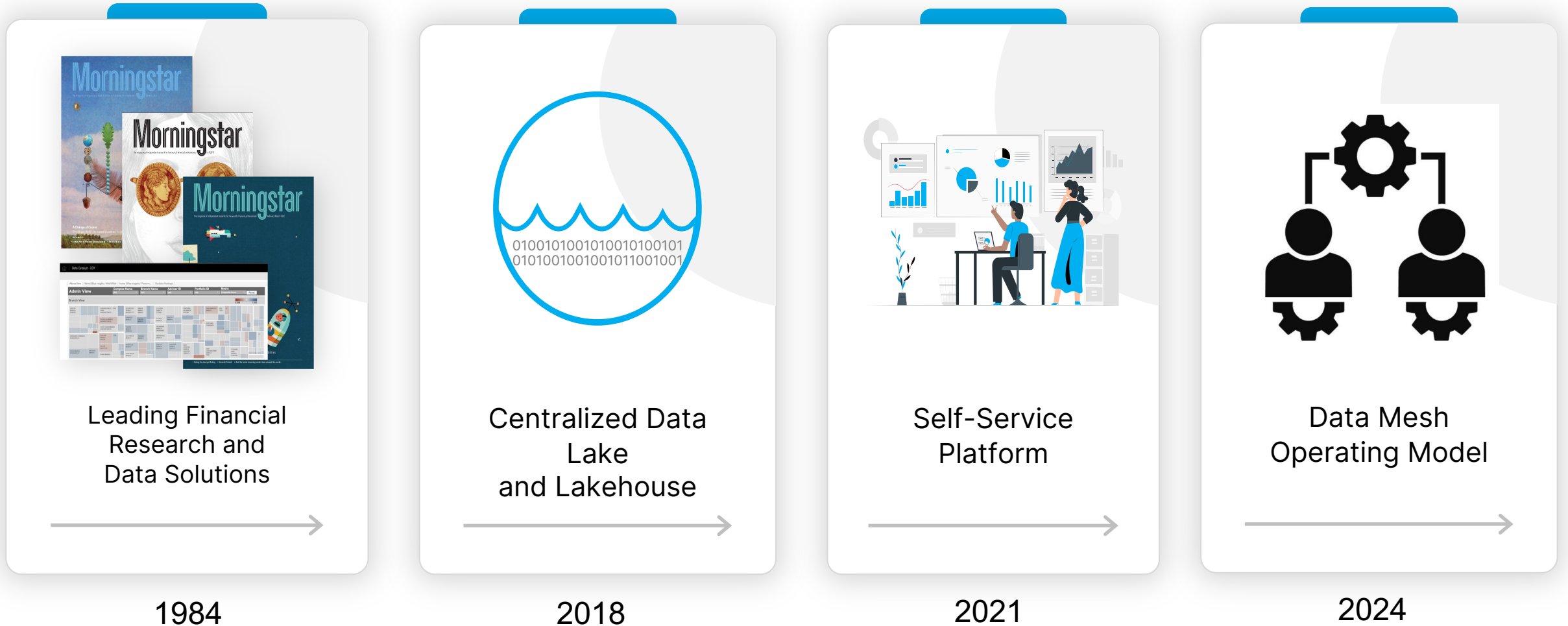
Hueler and Morningstar- a siloed operating model Before



Hopeful improvements example: Moving to a collaborative operating model



Democratized Data Mesh Journey



Thank you!

We are looking to collaborate and exchange ideas:



Annie Homsy-Harris



Anusha Dwivedula