The CDO's Guide to Implementing Gen AI for Analytics

Using a Data Fabric to Accelerate Time to Value

CDOIQ Symposium

July 2024

Promethium

Gartner
COOL VENDOR
2024

Kaycee Lai has a proven track record bringing new, disruptive technologies such as data fabric, data catalogs, data de-duplication, data virtualization, and hyper-converged infrastructure to market. With nearly 20 years of experience in the technology industry, Kaycee has led global operations & product management for both startups and Fortune 500 companies.

DELLEMC **vm**ware® (X) Delphix Waterline Data

Hitachi Vantara Promethium

Kaycee Lai
Founder, Promethium

Larry Llama

Why is GenAI Hard? Can't We Just Add a LLM to Our Data Source?

Data Sources

IBM DB2

snowflake

ORACLE

salesforce

Azure Data Lake

SAP HANA

HIVE

amazon REDSHIFT

Other data sources

LLMs still need help with these important steps

- Search for data across sources
- Decide between items with same name
- Verify veracity/usefulness of data
- Know when to combine across different sources

Gen AI / LLMs

ChatGPT

LLaMA by Meta
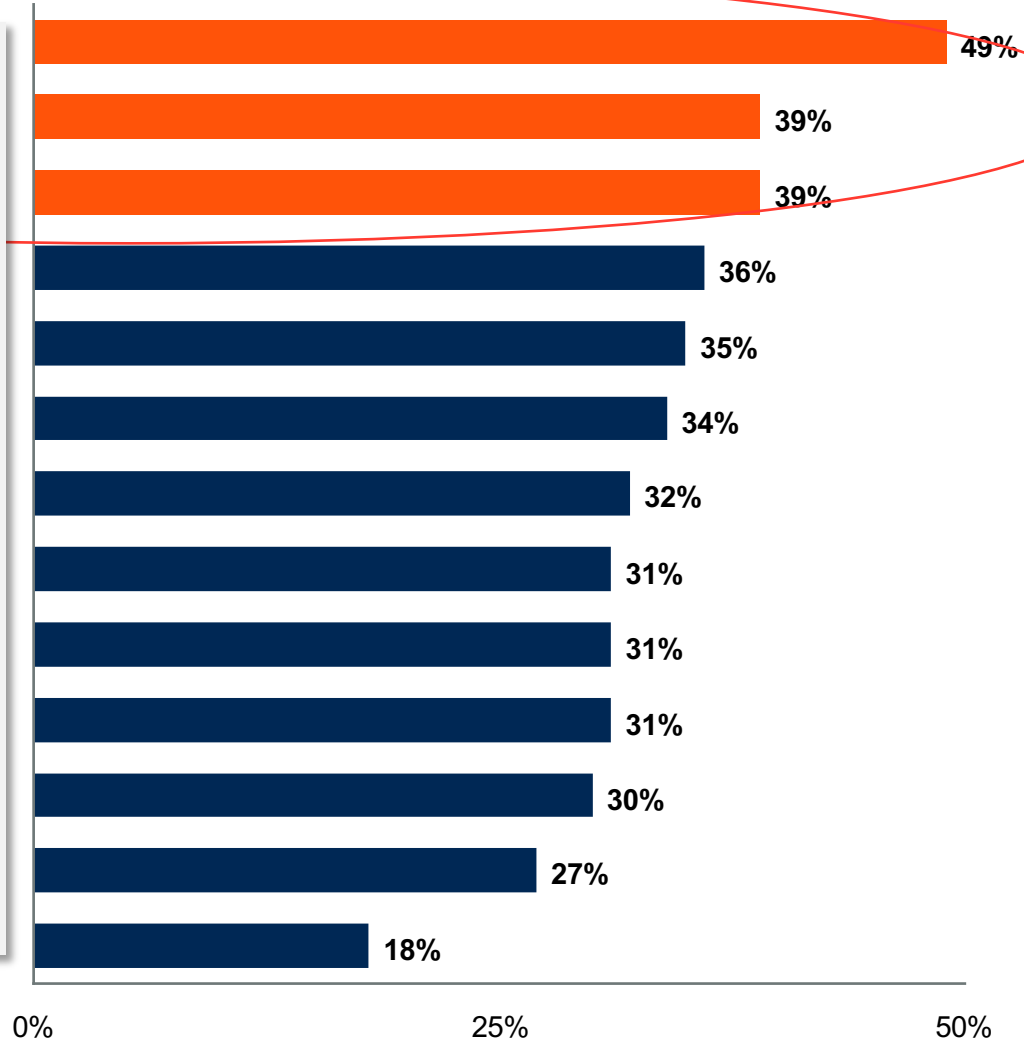
ANTHROP\C

Hugging Face

PaLM 2

Other LLMs

Gartner" Top Implementation Challenges for GenAI

- Finding value-add use cases is NOT the problem.

- The problems are:

  1. **How do I know the data is right and reliable?**

  2. **How to ensure the LLM works across all existing data sources?**

 3. **How can I make sure I'm not exposing my company to unnecessary risk**

Chart (bar values, top to bottom):
- 49%
- 39%
- 39%
- 36%
- 35%
- 34%
- 32%
- 31%
- 31%
- 31%
- 30%
- 27%
- 18%

X-axis: 0%, 25%, 50%

Promethium

CDOIQ Symposium

Addressing Key Challenges to Accelerate GenAI Projects into Production

1. Relevancy & Accuracy
Provide relevant context to consistently produce high quality responses to AI-driven questions
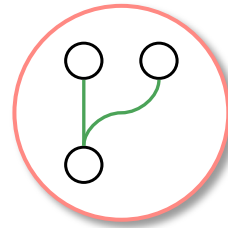
2. Data integration challenges
Data is fragmented across both new and legacy source systems in different formats & locations
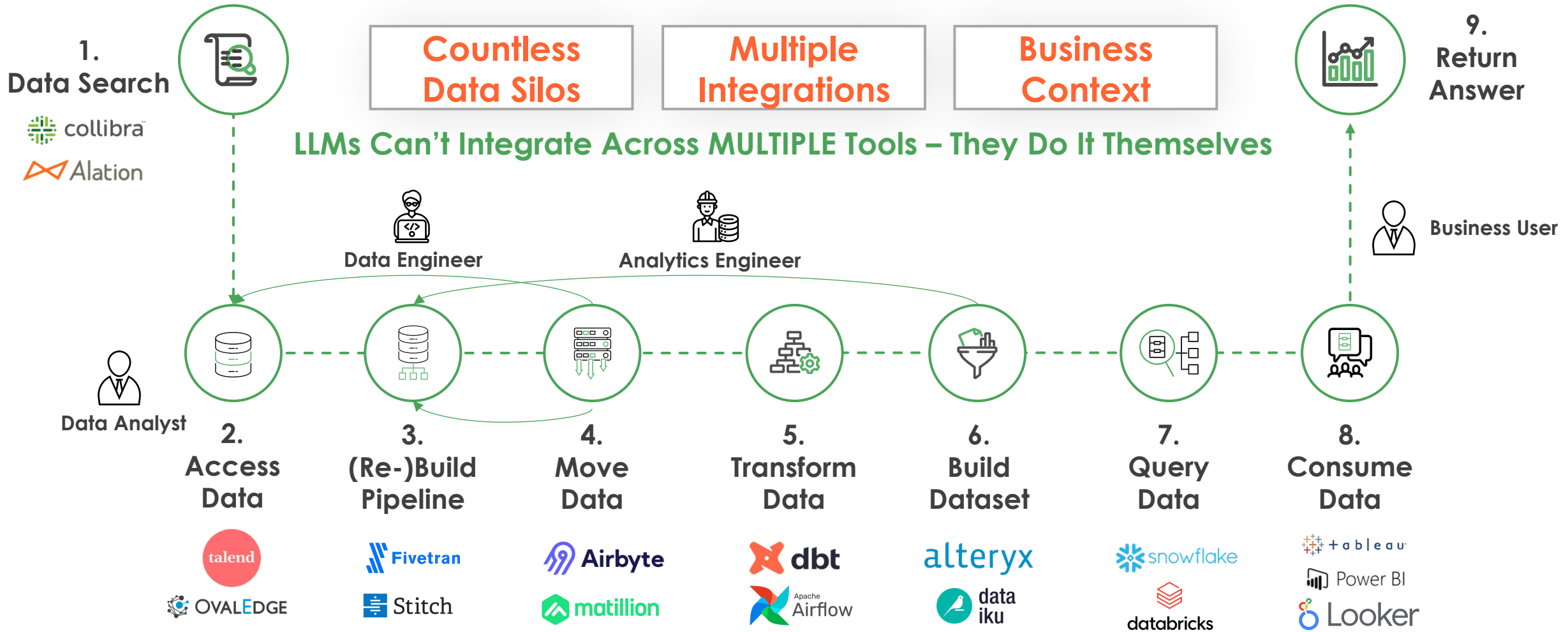
3. Security & Governance
Prevent sharing of proprietary data to train LLMs and ensure enterprise-grade access control

Challenge 1: Accuracy & Relevancy:
A lot happens today in siloed steps and products = Impossible Task for LLM.

1. Data Search

collibra
Alation

Countless Data Silos

Multiple Integrations

Business Context

9. Return Answer

LLMs Can't Integrate Across MULTIPLE Tools – They Do It Themselves

Data Engineer

Analytics Engineer

Business User

Data Analyst

2. Access Data

talend
OVALEDGE

3. (Re-)Build Pipeline

Fivetran
Stitch

4. Move Data

Airbyte
matillion

5. Transform Data

dbt
Apache Airflow

6. Build Dataset

alteryx
dataiku

7. Query Data

snowflake
databricks

8. Consume Data

tableau
Power BI
Looker

- A lot of products were built pre-AI or even pre-cloud.

- A lot of products lack REST APIs

- A lot of products w/ custom logic

- These products are not going anywhere

- Many sources can still be on-premise while the LLM may be in the cloud or vice-versa

Your LLMs

Using a 1:1 Relationship of LLM to Data Source Is NOT the Answer

salesforce **LLM1**

snowflake **LLM2**

Azure
Data Lake **LLM3**

databricks **LLM4**

amazon
REDSHIFT **LLM5**

How Does Self-Service Get Easier?

Each LLM trained on different data

Different way to prompt

Isolated knowledge & data access

Data to answers may not be in real-time

Challenge 3: Data Protection & Privacy

Send data to a Public LLM

Load all data to your Data Lake/Warehouse which has a LLM
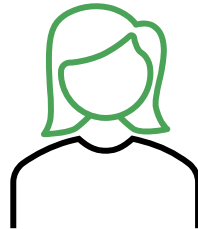

VS

Access Controls Must Also Apply to LLMs
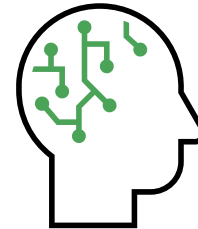
CDOIQ Symposium

"SLA" has changed

Data **delivery needs to be real time** versus weeks / months

"Persona" has changed

Business users want to **access and consume data conversationally**

Single-user workflows

Streamlined workflows require **unified data stack**

Gen AI requires a new approach to the modern data stack that is quick, accessible, and secure

Solution

...There is Hope

Promethium

Data Fabric & Active Metadata Bridge Gaps to Power GenAI

Data Sources

IBM DB2 · snowflake

ORACLE

Azure Data Lake · salesforce

SAP HANA · HIVE

amazon REDSHIFT

Structured data

Semi-structured data

Unstructured data

Promethium

Gen AI / LLMs

ChatGPT

LLaMA by Meta

ANTHROP\C

Hugging Face

PaLM 2

+ others

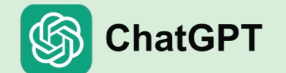**Provide <u>real-time data access</u> with <u>relevant context</u> for <u>endorsed</u> data enabling Generative AI via an AI-powered solutions**

A Data Fabric Leverages a Flexible Approach…

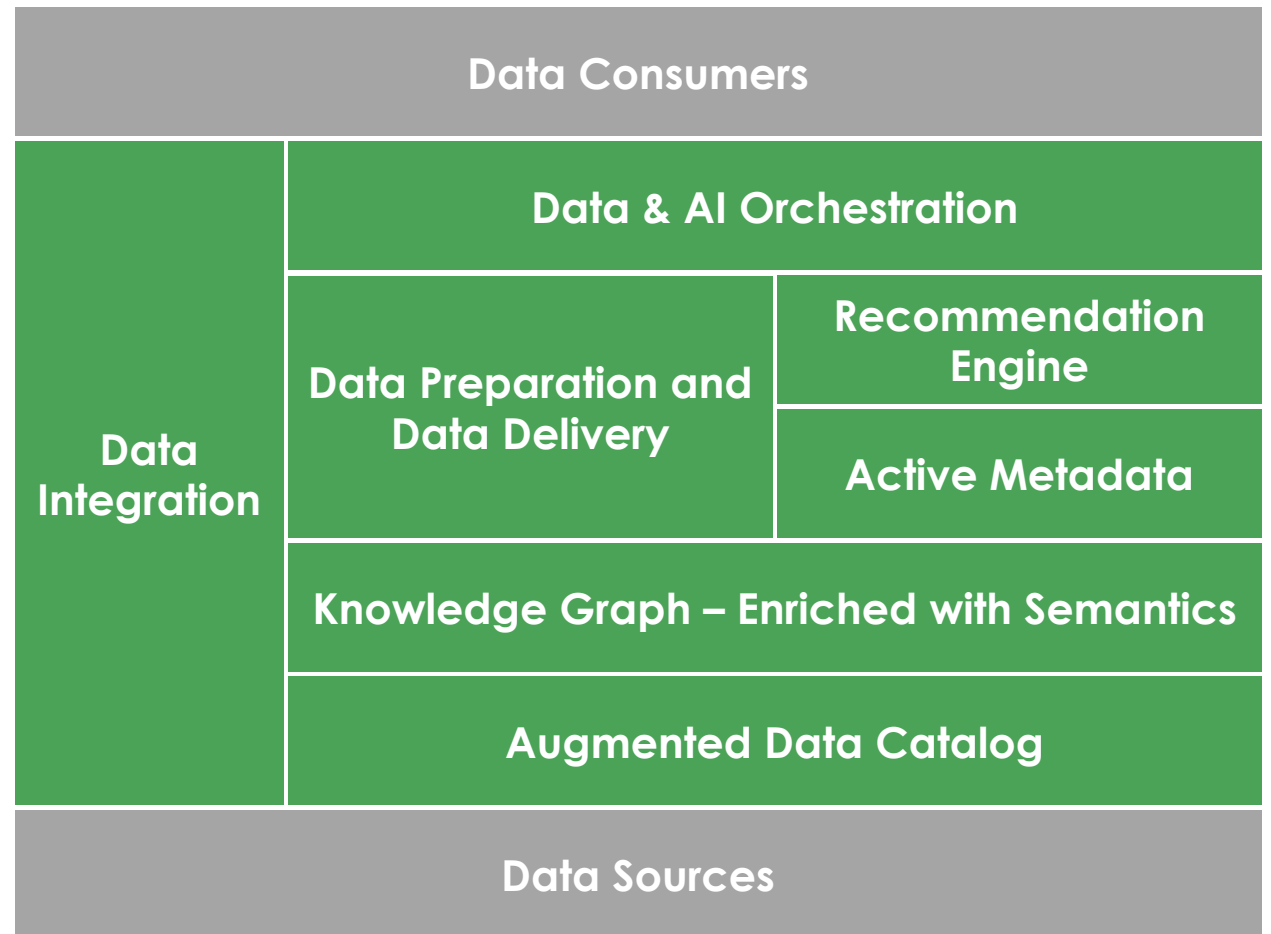*"Data Fabric overturns the dominant approach to data management which is "build to suit" for data and use cases and replaces it with "observe and leverage."*

"Promethium provides a **platform for creating a data-fabric-style architecture, which also automates data product creation**.

"In particular, this product is a good match for organizations that do not want to use professional services and that are looking for quick time to value"
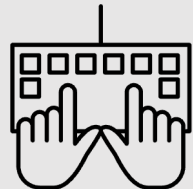
© Gartner

Gartner
COOL
VENDOR
2024
™

Data Consumers		
Data Integration	Data & AI Orchestration	
	Data Preparation and Data Delivery	Recommendation Engine
		Active Metadata
	Knowledge Graph – Enriched with Semantics	
	Augmented Data Catalog	
Data Sources		

Promethium

CREATING DATA

DATA FABRIC

1

Where Data is Created

Data is Created in tools across your organization

These **systems cover a range of functions** that help you run your business and manage the grid. Things like…
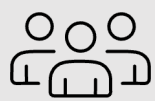
Finances
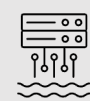
Assets

Work

Customer

Tools for Moving Data IF needed

Where we access it

The **best method to access this data is directly** in the tools where it is created, but sometimes you need to move it. This could be because the tool may be old or isn't designed for lots of people accessing the data. We move data to…

Data stores

Data lakes

2

8 The Fabric is also the place where we get **full visibility of all our data and products** and gives us the ability to **govern connections and products through contracts**

6 The Fabric enables a **common data language**
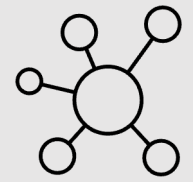
7 The Fabric **provides security and protection** for our Data and Products to **keep us Safe!**

**3** The first thing a Data Fabric does is **connect to all our Data** allowing us to access it

But **connection isn't enough!** The Data Fabric helps us to **understand our data** by…

…**mapping** all our data, creating a **catalogue** about our data and providing an easy way to search and **discover** it!

4 The next thing a Data Fabric does is provide **a place to build** what we need to do with our connected data
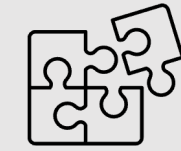
What we build is important! It is how we **add value and insight** to the data and make it easy to use through…

…**creating collections of data** known as Data Products, **applying intelligence** to improve analysis and **cataloging** products for discovery!

Finally, we **Market our Products** via a Portal **5**

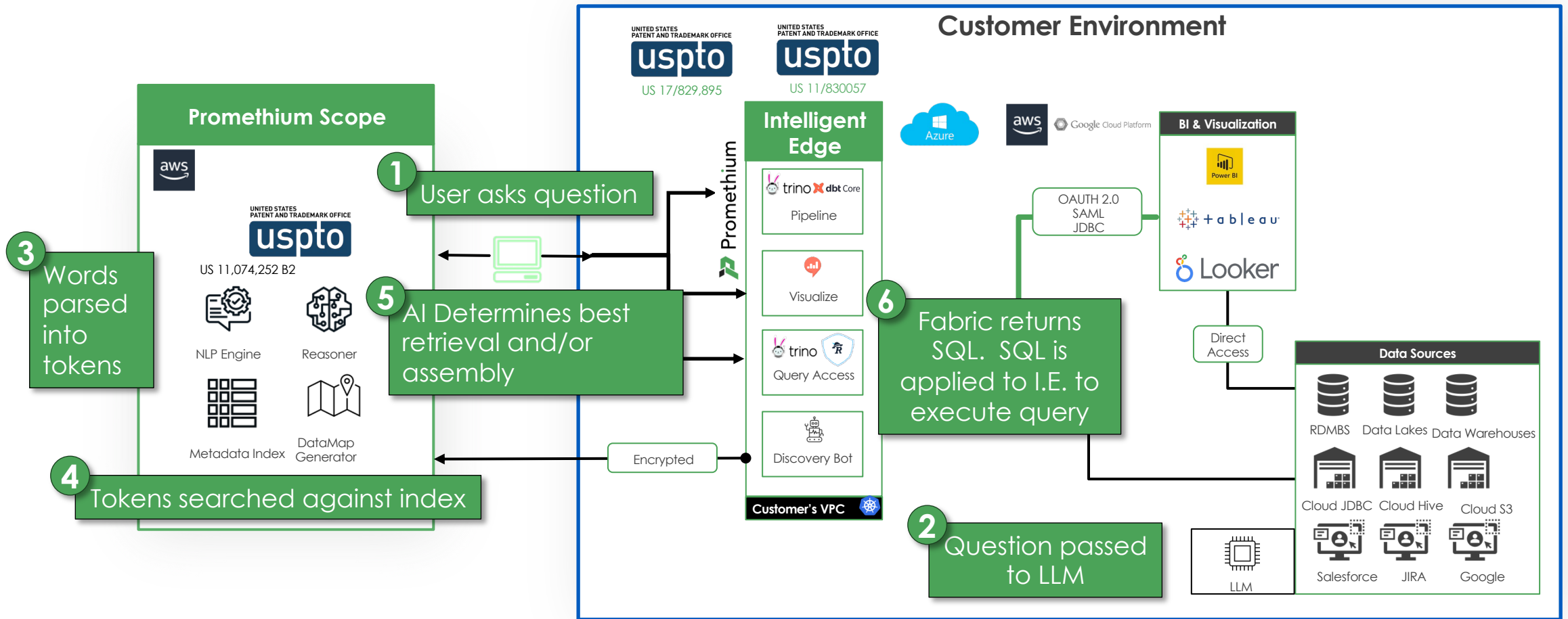It 'opens' our Products **internally and externally**…

…providing **simple search** and **easy connection** to data insights

DISCOVER & ACCESS

ADDING VALUE: BUILD & SHARE

Architecting Security, Performance and Automation for GenAI

Promethium Scope

aws

UNITED STATES PATENT AND TRADEMARK OFFICE
uspto
US 11,074,252 B2

NLP Engine Reasoner

Metadata Index DataMap Generator

1 User asks question

3 Words parsed into tokens

5 AI Determines best retrieval and/or assembly

4 Tokens searched against index

Customer Environment

UNITED STATES PATENT AND TRADEMARK OFFICE
uspto
US 17/829,895

UNITED STATES PATENT AND TRADEMARK OFFICE
uspto
US 11/830057

Promethium

Intelligent Edge

trino dbt Core
Pipeline

Visualize

trino Query Access

Discovery Bot

Customer's VPC

Encrypted

Azure aws Google Cloud Platform

OAUTH 2.0 SAML JDBC

BI & Visualization
Power BI
tableau
Looker

Direct Access

6 Fabric returns SQL. SQL is applied to I.E. to execute query

Data Sources
RDMBS Data Lakes Data Warehouses
Cloud JDBC Cloud Hive Cloud S3
Salesforce JIRA Google

LLM

2 Question passed to LLM

Success Story: F500 Blood Testing Company

- **Increase visibility and insights**, particularly around patient care and billing
- **Enable Generative AI** across multiple use cases
- **Empower self-service analytics** to respond to patient and physicians needs in real time

Mark Clare
VP & CDO

"With quickly advancing data and AI technologies, if I designed just for today it could be obsolete before I fully deliver….."

"I need flexibility and extensibility in that core to adopt emerging technical capabilities as data products on the fabric …"

"People want to have a conversational approach to access the data they need to get answers."

Promethium

Data Fabric Offloads Prompt Engineering and Introduces Enterprise Context

■ Automated or augmented processes

■ Human action or machine learning retraining required

■ Already exists

⚙ Machine-driven analysis creates new or inferred metadata for further processing

⚙ Human is creating something new that the machine cannot yet learn

⚙ Process existing metadata

Traditional Data Management Starts With Human Coding Patterns

Fabric Recognizes Enterprise Patterns and Tests LLM Proposals

Fabric Learns from Production

Fabric Observation

Machine Coding Patterns → Generate → LLM Proposal

ML Model · Patterns · Content

Out-of-the-Box LLM

Formalized and Literal Code Languages

Code Versions (Similar to Dialects)

Over 80 Years and Thousands of Language Uses

Coding Patterns and Practices

Design and Use Prompts

Test LLM Code
- Run Process
- Get Output
- Runtime Metadata

- Production Runs
- Existing Outputs
- Runtime Metadata

Identify Exception < = > and Alignment

Retrain or Custom Model | Recommend | Challenge | Machine Dialog | Validate

Fabric Isolates and Protects Enterprise IP

Source: Gartner
796172_C

Gartner®

"The relationship between active metadata and data fabric is mandatory for assuring validity and transparency in data management operations, enabling LLMs to become the preferred interface to enterprise data."

"To the data fabric, the LLM is just another developer. To the fabric, there is no difference between human agent and a machine agent."

"The combination of the data fabric and GenAI capabilities promises that manual data management tasks will be automated or AI assisted."

What If..There IS ONE Logical View of ALL of Your Data
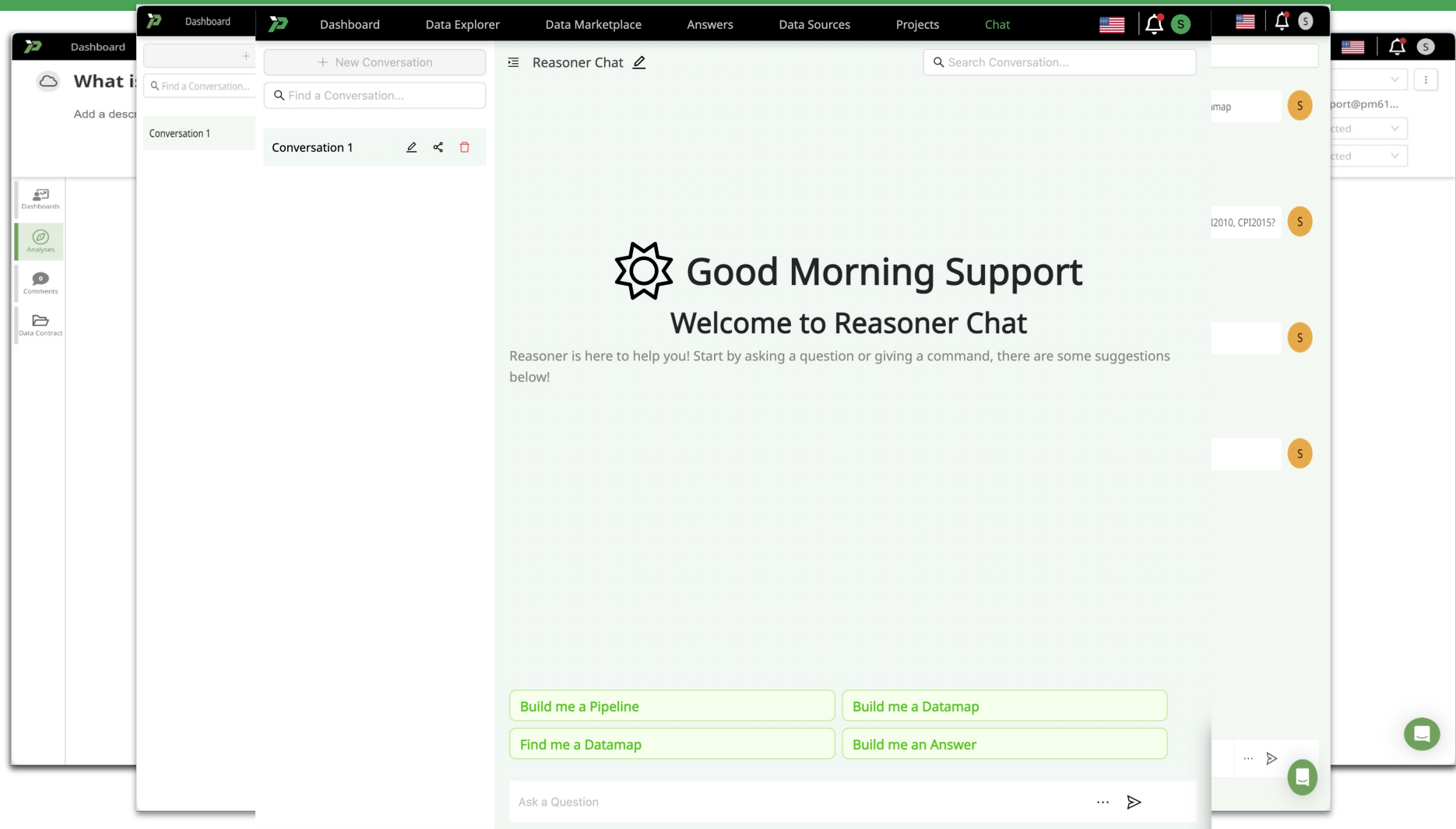## WITHOUT Moving / Duplicating Your Data

- Automatically crawl and catalog over 200 data sources
- Automatically annotate and add AI-powered descriptions to data
- Use natural language prompts to discover your data
- Preview and explore your data **without moving or storing data**
- Search and inspect all data from 1 browser
- **Provide 1 SINGLE view to the LLM**

1 Promethium's Data Fabric creates a unified logical view from all sources

2 LLM talks to Data Fabric NOT databases, data lakes, data warehouses

3 No data is sent to or stored in the LLM, ensuring security & governance

Powered a prompt-based approach to data management

Our unique approach
Building data products with AI & Active Metadata
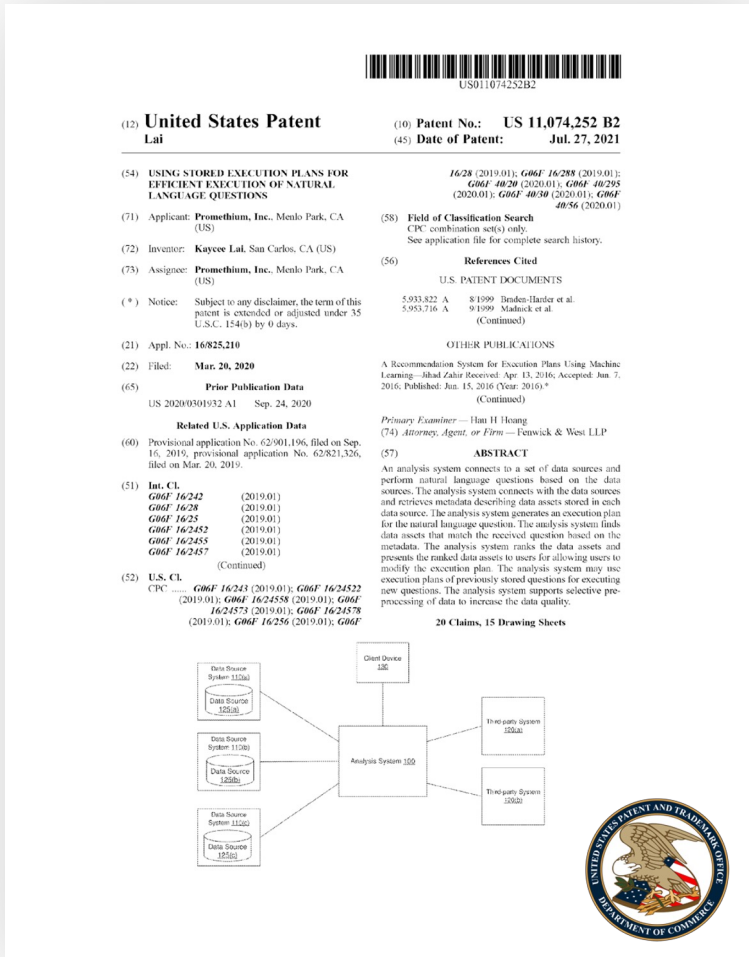
Faster Insights
Eliminate wait time
for the business

Answer questions
in minutes via
self-service

Boost Productivity
Automate
repetitive tasks.
Reduce complexity

Collaborate in
real time Re-use
vs Re-do

Modern
Future-proof workflow
from changing new
data sources & tools

1 solution vs 6+
80% less costs

5.
Return Answer

1.
Ask Question

2.
Generate Datasets or Query

3.
Query Data

4.
Consume Data

To automatically build new data products from scratch

Q: "What is the revenue subscription by vendor by package from 2013 to 2019?"

Tokens Found in the question are:
- date
- revenue
- package
- vendor
- subscription

REASONER identified and grouped them in the below categories:
Measure Tokens: revenue,subscription
Filter Tokens: date
Order By Tokens: N/A
Group By Tokens: vendor,package

What tables have been shortlisted by REASONER and Why?

TABLE_NAME	IDENTIFIED AS	MATCHED TABLE NAME	MATCHED TABLE TAGS	MATCHED FIELD NAMES	MATCHED FIELD TAGS	MATCHED TERMS	ENDORSED	IS PRIMARY TABLE	DATAMAP REF
postgresql.customer_demo.fact_subscription_activity	join, select (measure) TABLE(S)	Yes	revenue, subscription			revenue, subscription	Yes	Yes	–
redshift.customer_demo.dim_agent	group_by TABLE(S)	No		agent_vendor	vendor	vendor	Yes	No	–
oracle.CUSTOMER_DEMO.MASTER_PLAN_TABLE_WDATES	group_by, join, select (measure) TABLE(S)	Yes		package, startdate, enddate	revenue	date, revenue, package	Yes	No	–

Join selection by REASONER

COLUMN_1	COLUMN_2	REASON
oracle.CUSTOMER_DEMO.MASTER_PLAN_TABLE_WDATES.PLAN_ID	postgresql.customer_demo.fact_subscription_activity.PLAN_ID	Based on existing join from some datamap
postgresql.customer_demo.fact_subscription_activity.AGN_KEY	redshift.customer_demo.dim_agent.agent_record_key	Column { agent_record_key } contains a tag { agn_key }

Leverage AI to build datasets, with visual lineage and explanation from the AI.

Success Story

- **Enable GenAI**, to make complex SAP data easy to find, understand and access
- **Accelerate queries** for large datasets across multiple sources (reduce times by 98.6%)
- **Empower self-service analytics** to identify supply chain issues with natural language to avoid financial penalties

"I'm in a fast-paced business. Even a 15 min delay may not work for me. With Promethium, we can answer questions on key issues around supply chain in near real-time so we can make changes that save $$$. The fact that we get a single view of all our data and we don't have to move / copy data into one source saves time & money."

Jeremy Vance
VP Information Technologies
Hostess Brands

Promethium Data Fabric: Designed for GenAI
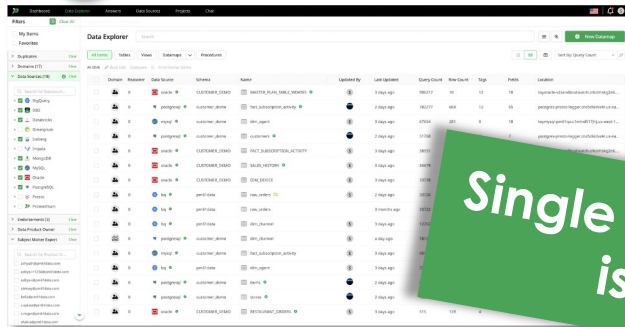## With Security, Relevance, Accuracy Baked in for a SINGLE Workflow

Promethium

Single No-Code Solution to Build Dataset | **Complete Workflow in Minutes** | **No Need to Wait First**
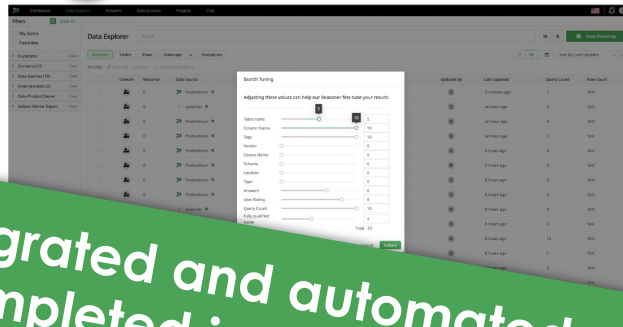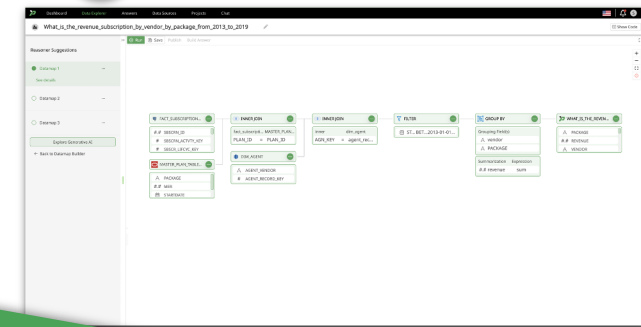
1 Discover

2 Active Metadata

3 Auto Assemble
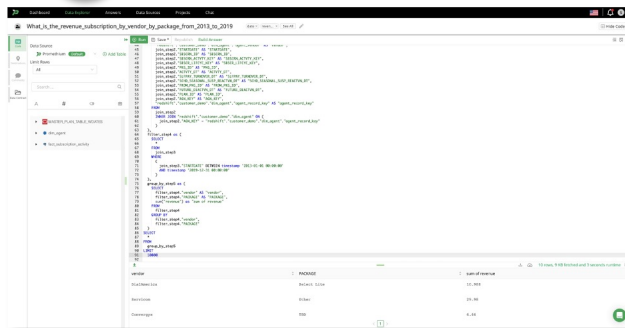
Single integrated and automated workflow that is completed in a matter of seconds

4 Query / Safe Access

5 Validate / Co...

Govern

Solution: Providing Necessary Context for Accurate GenAI

Downstream Data Products

Lineage

Historical joins / Semantic models

Promethium

Crowdsourcing

Domain / Team / Department

Frequency of Use
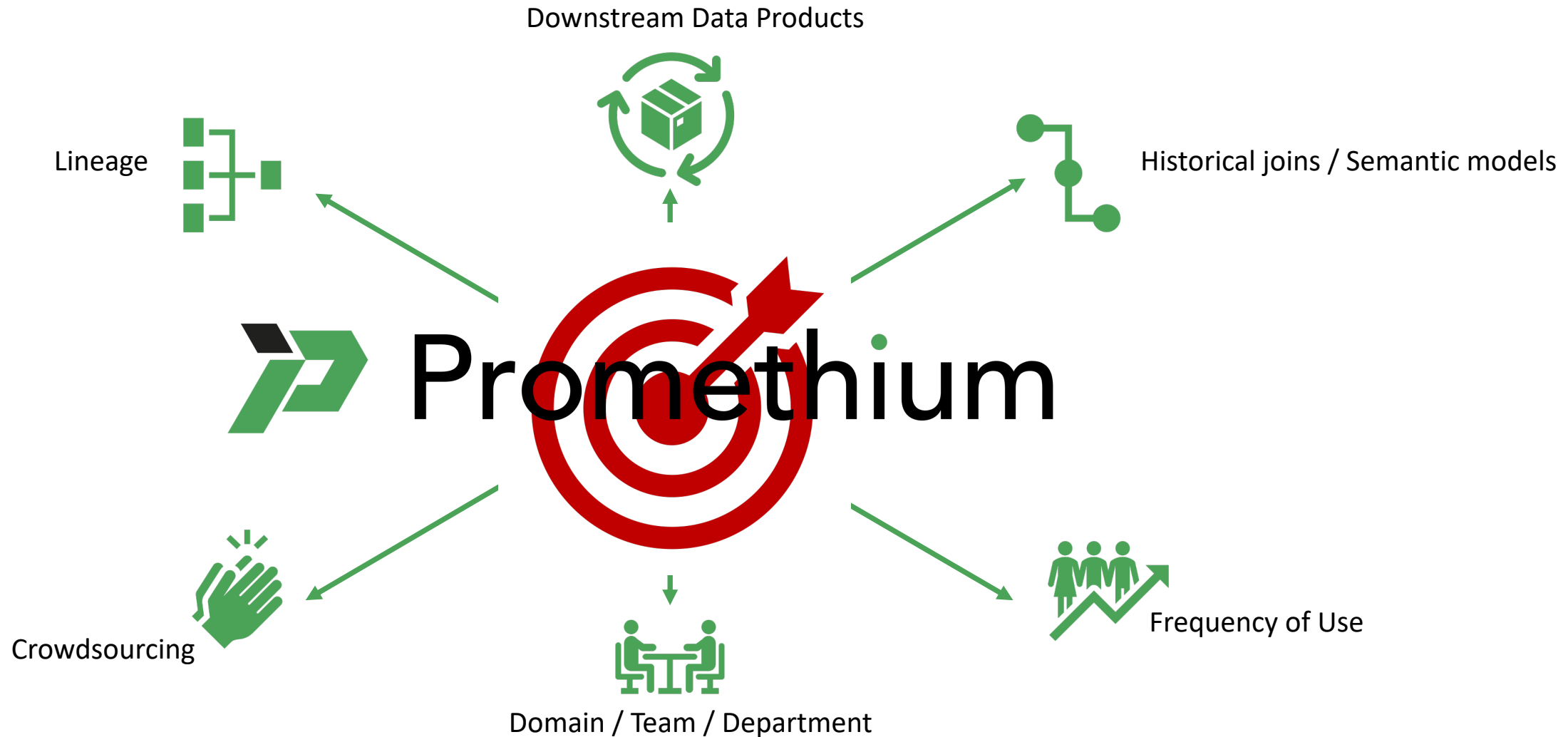
Thank You

Kaycee Lai
Founder

Promethium