

The background features a dark blue gradient with abstract, glowing shapes in shades of purple and pink. Two thin, light blue lines intersect to form a large 'A' shape. The text is positioned on the left side of the image.

AWS re:Invent

DECEMBER 2 - 6, 2024 | LAS VEGAS, NV

ANT207

Empower your data journey with Amazon DataZone data lineage

Priya Tiruthani

Sr. Product Manager
Amazon DataZone
AWS

Harel Shein

OpenLineage TSC
Sr. Eng. Manager
Datadog

Rob Malowney

Group Product Manager,
Enterprise Data & Governance
San Diego Gas & Electric

Leonardo Gomez

Prin. Solution Architect
AWS Analytics
AWS

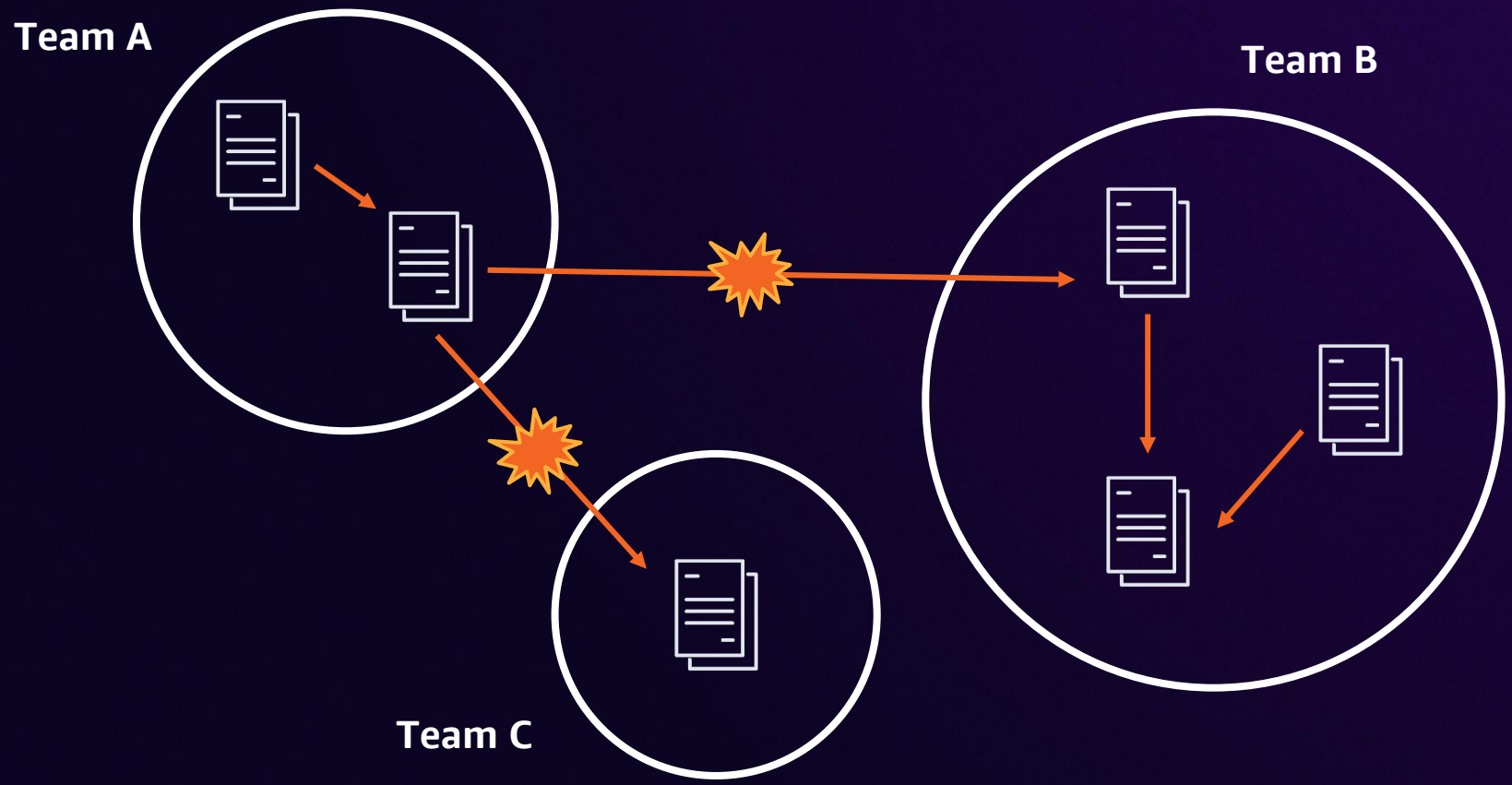


Agenda

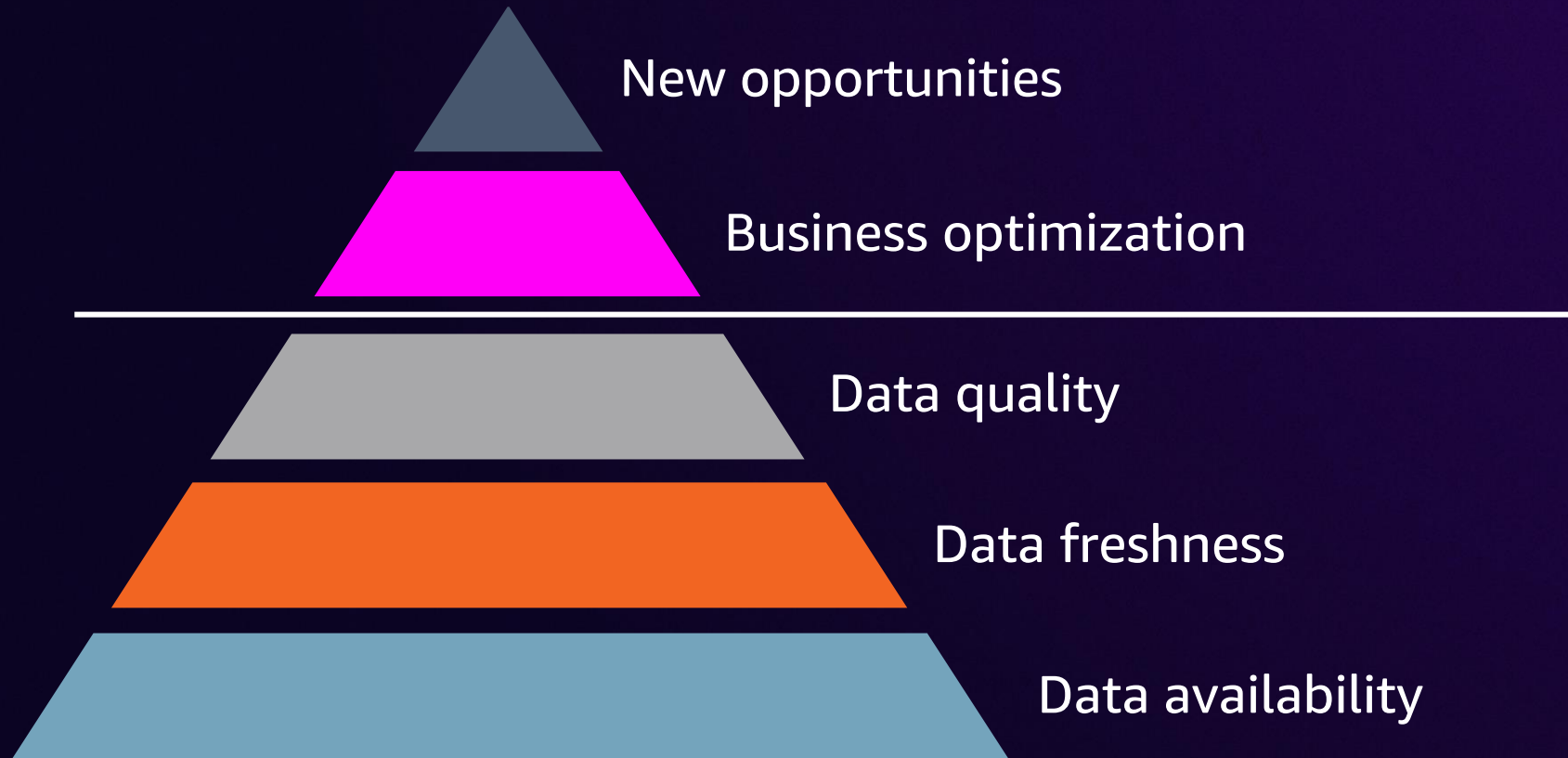
- Why lineage and what are customers looking for?
- Introduction to data lineage in Amazon DataZone
- Why OpenLineage and why does it matter?
- Customer journey on Amazon DataZone data lineage
- Q&A

Why do we need data lineage?

Building a healthy data ecosystem



~~Maslow's~~ Data hierarchy of needs



The possibilities are endless

Dependency tracing
Root cause identification
Issue prioritization
Impact mapping
Precision backfills
Anomaly detection
Change management
Historical analysis
Compliance



What some vendors say about lineage

Fully automated

Real-time

End-to-end

360° visibility

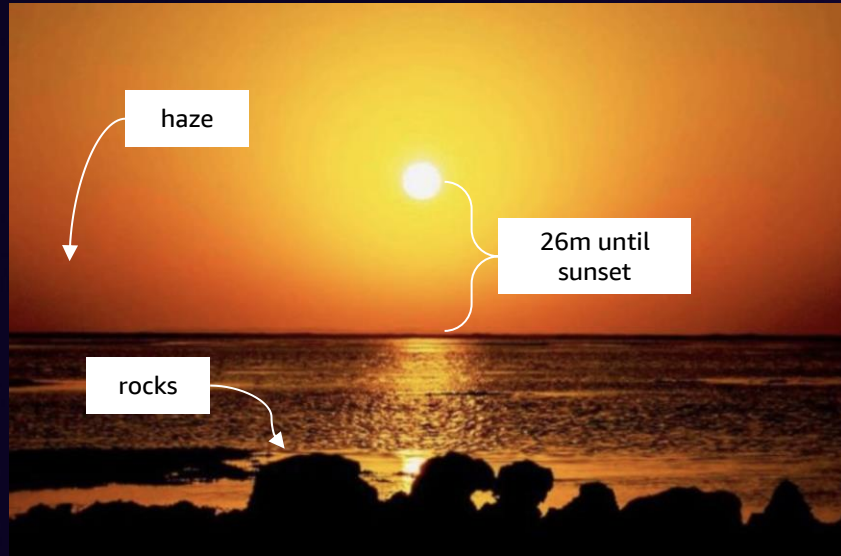
Easy

AI/ML-powered

OpenLineage design principles



The best time to collect metadata



You can try to infer the date and location of an image after the fact . . .



. . . or you can capture it when the image is originally created

OpenLineage

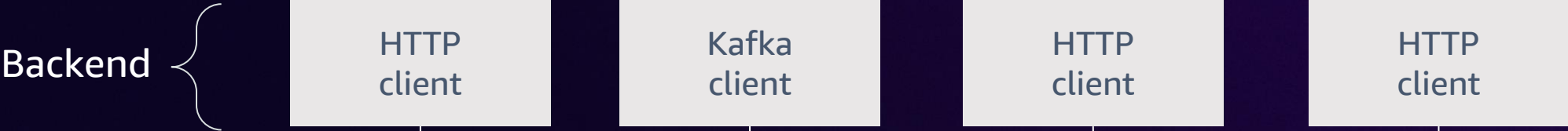
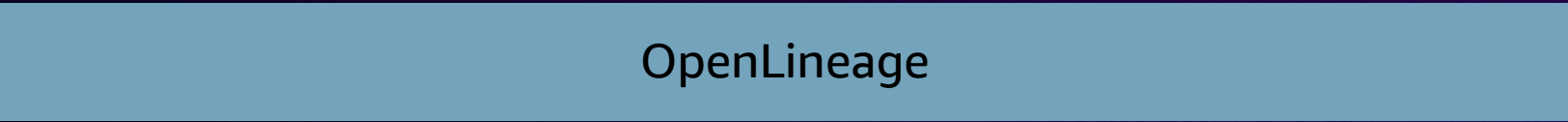
Mission

To define an **open standard** for the collection of lineage metadata from pipelines **as they are running**

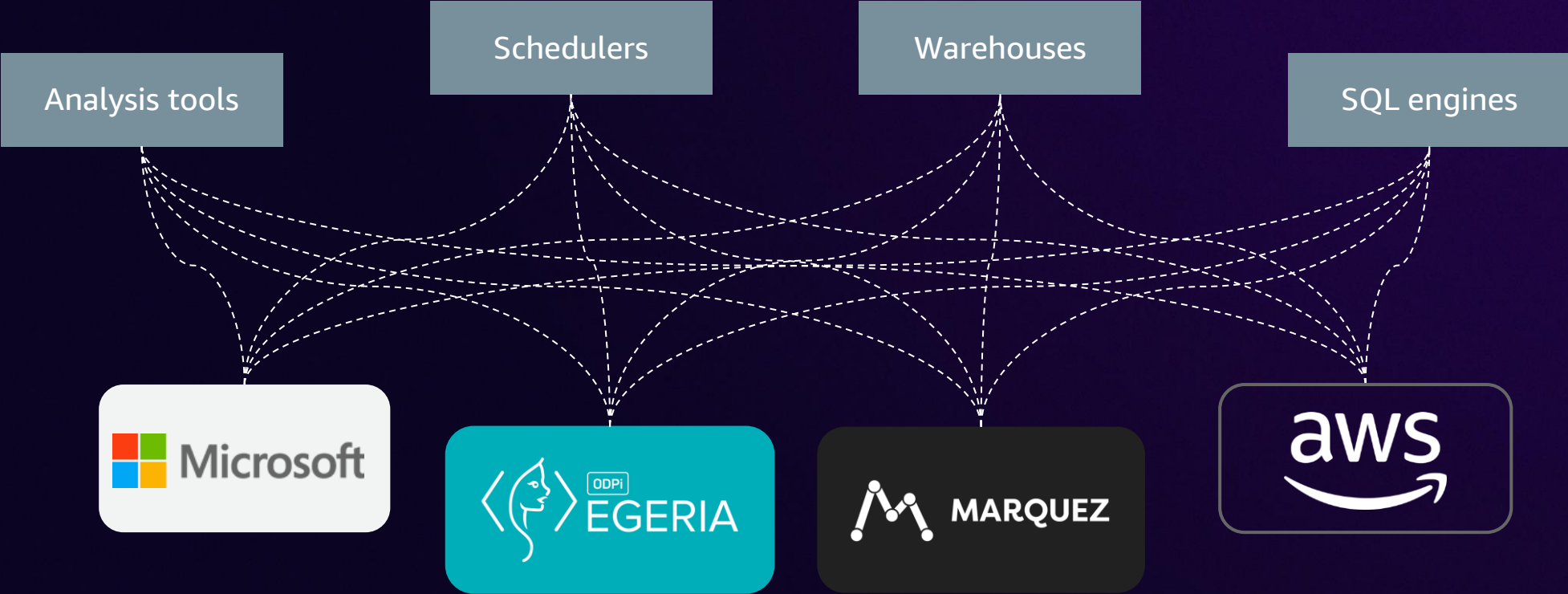
(This is evolving)



Where OpenLineage fits



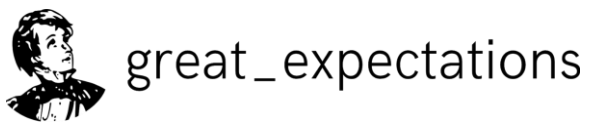
Before OpenLineage



With OpenLineage

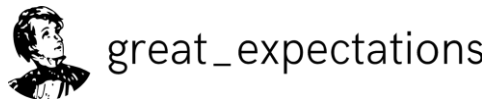


OpenLineage contributors



OpenLineage integrations

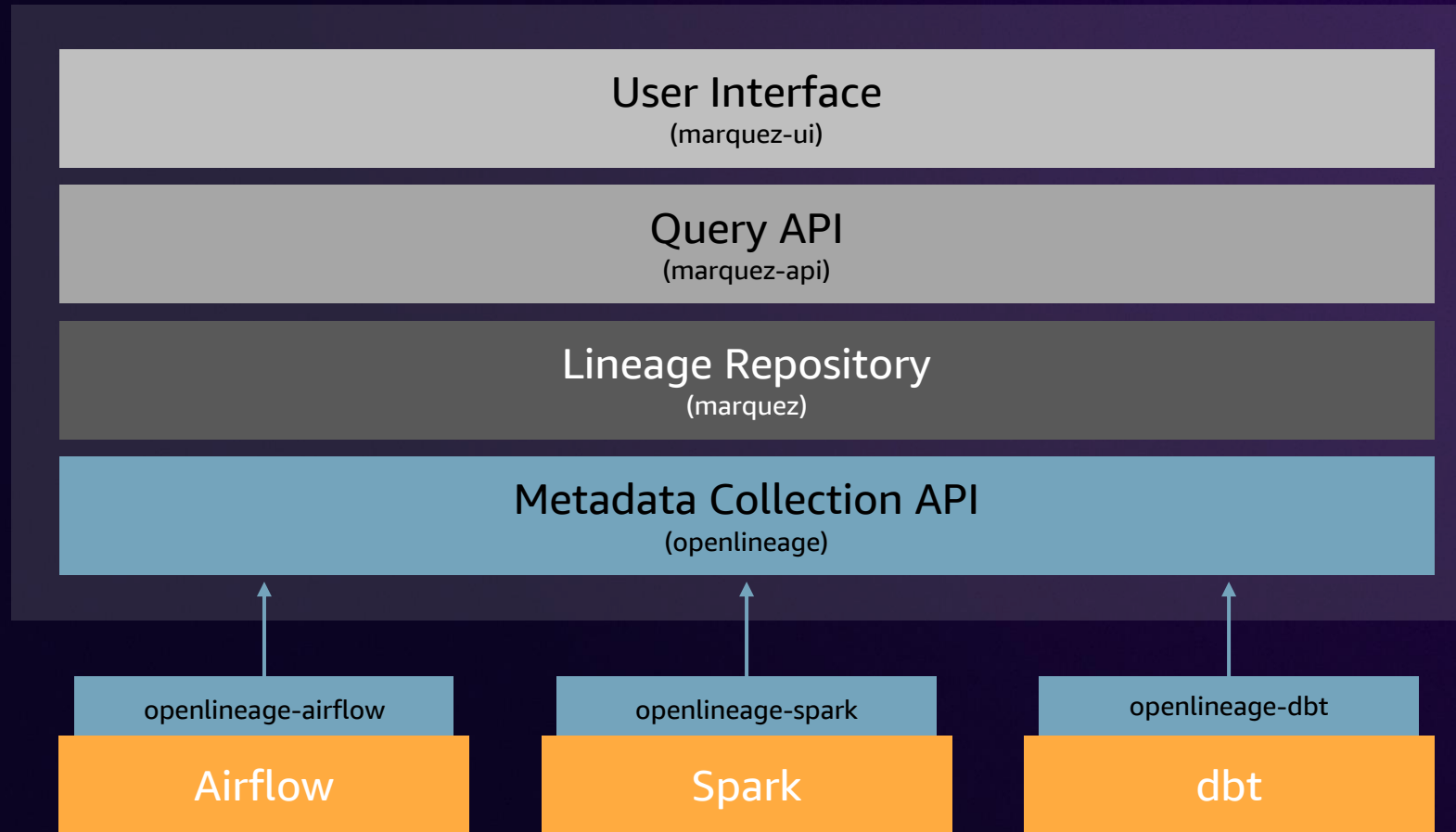
Metadata producers



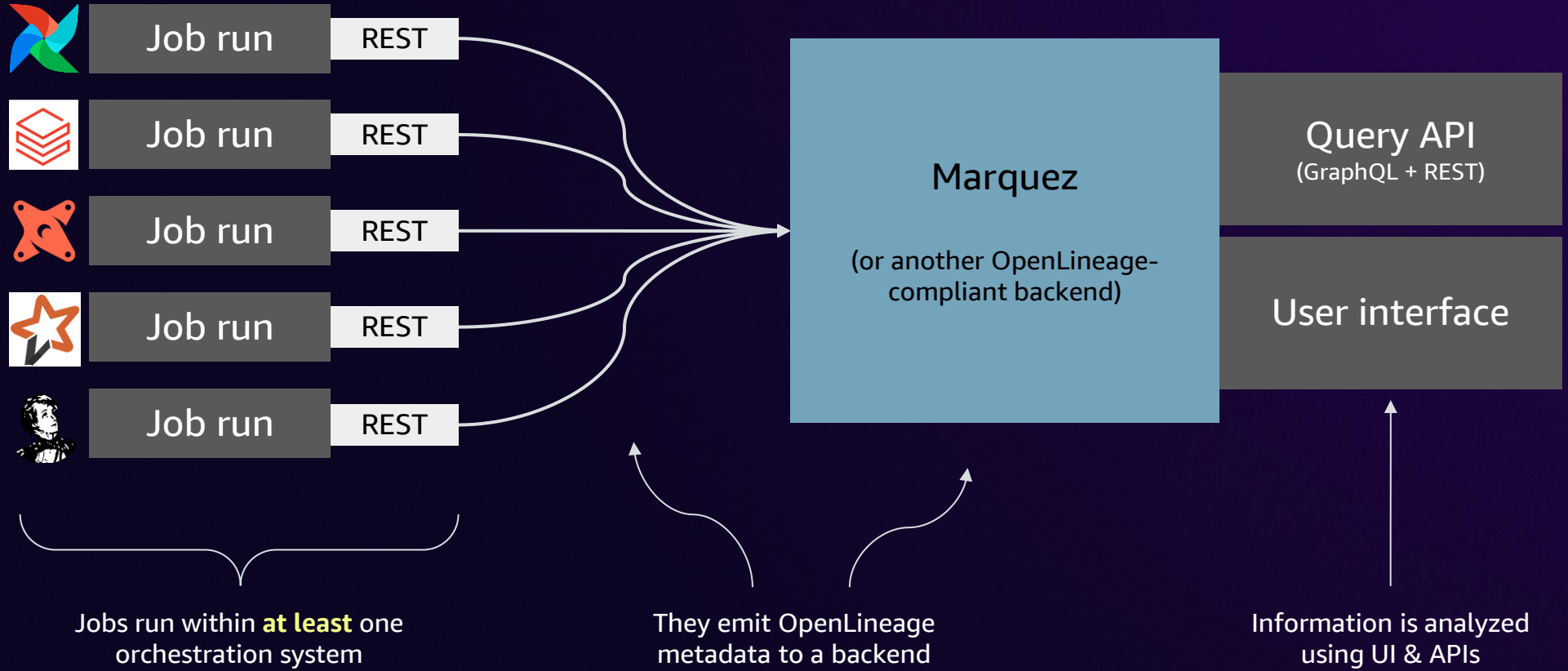
Metadata consumers



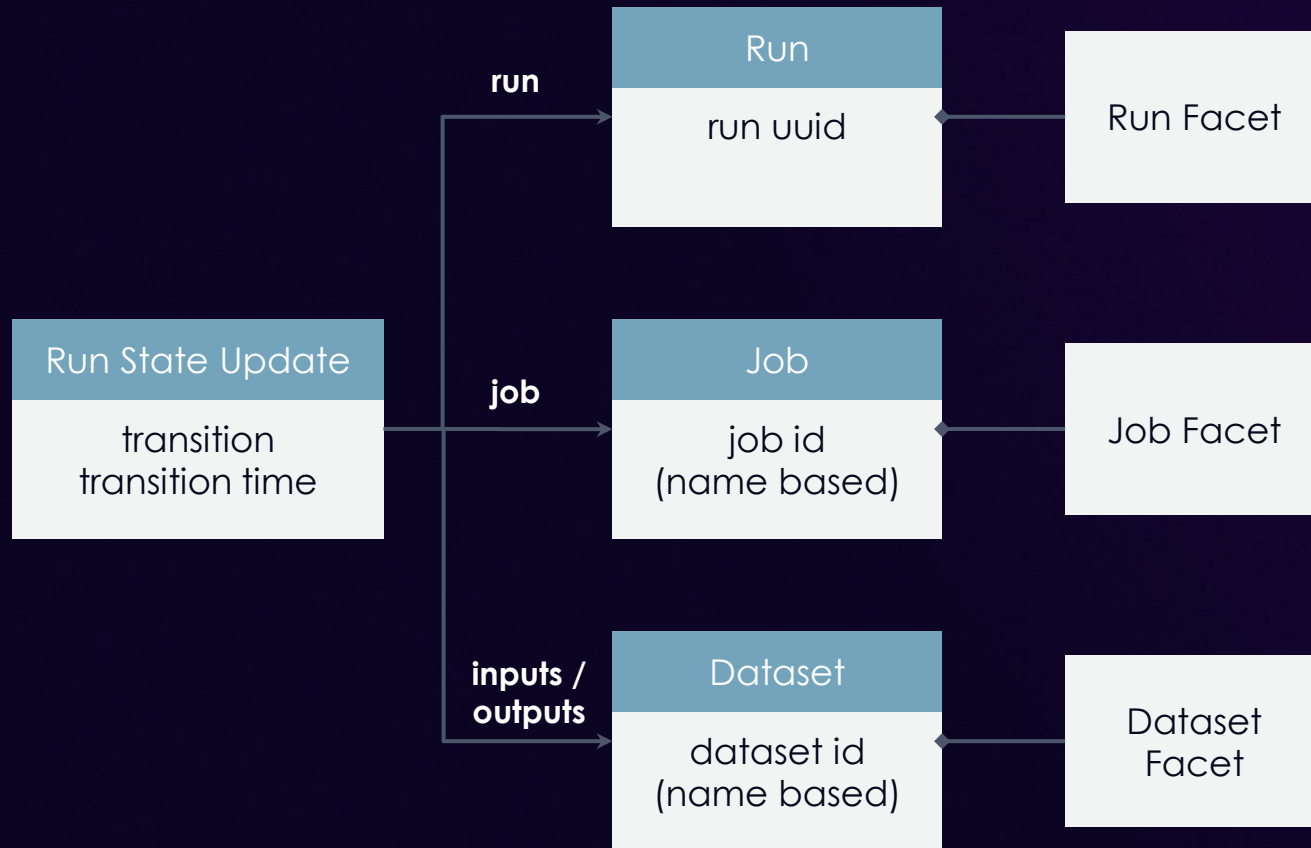
The OpenLineage (reference) stack



OpenLineage uses a "push" model



Data model



Built around core entities:
Datasets, Jobs, and Runs

Defined as a JSON
Schema spec

Consistent naming for:
Jobs (*scheduler.job.task*)
Datasets
(*instance.schema.table*)

Lifecycle of a job run



Extending the model with facets

Facets are atomic pieces of metadata attached to core entities.

Self-documenting

Facets can be given unique, memorable names

Familiar

Facets are defined using JSON schema objects

Flexible

Facets can be attached to any core entity: Job, Dataset & Run

Scalable

Prefixes on names are used to establish discrete namespaces

Data lineage in Amazon DataZone



Themes we hear from customers

Graphical view to help users be more productive, make better decisions, and comply with regulations

Trust

Build trust to ensure data assets are matched to correct use cases

Impact analysis

Reduces the time required for change management and troubleshooting activities

Troubleshooting

When a data issue is reported, the issue source should be easily identified by the user diving deep

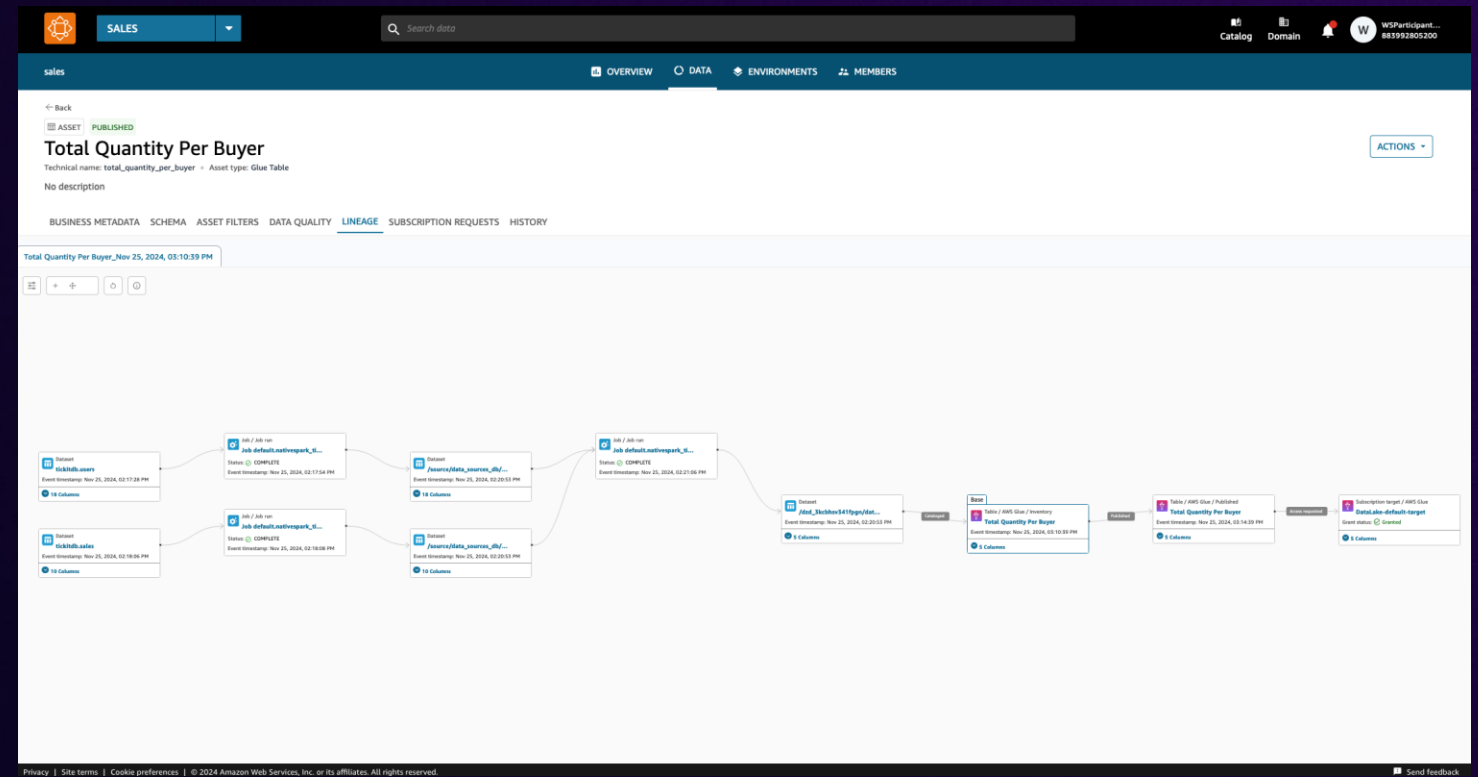
Governance

For audit purposes, provide scalable approach to see how data assets are used and accessed

Data lineage in Amazon DataZone

What's available in GA

- **Automated lineage** from AWS Glue and Amazon Redshift data sources including AWS Glue ETL and notebooks
- API support – **OpenLineage compatible**
- **Interactive visualization** with dataset and column lineage with versioning
- **Enhance with AI lineage using custom assets** to dive into AI models, dashboards, or other assets



Meet the marketing team!



Marina

Marketing analyst

I need to confirm the origin of a data asset to confidently use it in my analysis



Julia

Data engineer

I need to understand the impact of my work on dependent objects to avoid unintended changes



Julia

Data engineer

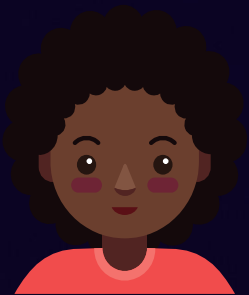
I need to investigate why a report is showing incorrect data and quickly fix what broke along the way



Susan

Administrator

I need to fulfill audit requests by tracing reporting figures back to sources and identifying transformations applied



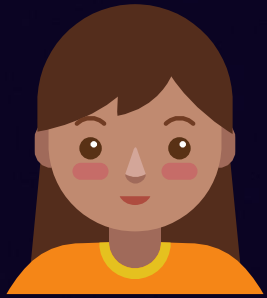
Marina
Marketing analyst

- Technical/business metadata of upstream connections
- Usage information of the data asset (i.e., used by 75 reports)
- Origin information across all connected data assets

Data origination

Marina begins by looking at the Lineage tab on the data asset detail page. From there she navigates upstream to understand data provenance.

The screenshot shows the AWS Glue console interface. At the top, there's a navigation bar with 'ADMIN', 'Search Assets', and user information 'frontend-user'. Below that, a secondary navigation bar shows 'Admin', 'OVERVIEW', 'DATA', 'ENVIRONMENTS', and 'MEMBERS'. The main content area is titled 'Inventory' and includes technical details: 'Technical name: inventory', 'Asset type: Glue Table', and 'This is the data asset for the Inventory table.' There are buttons for 'ACTIONS' and 'RE-PUBLISH ASSET', with a note: 'New revision is available to publish. Latest revision: 4'. Below this, there are tabs for 'BUSINESS METADATA', 'SCHEMA', 'ASSET FILTERS', 'DATA QUALITY', 'LINEAGE' (selected), 'SUBSCRIPTION REQUESTS', and 'HISTORY'. The 'LINEAGE' tab shows a flow diagram for 'Inventory_Jul 14, 2023, 05:49:26 PM'. The diagram consists of three columns of nodes. The first column has three 'Dataset' nodes: 'Inventory_us_east', 'Inventory_us_west', and 'Inventory_us_central'. The second column has three 'Job / Job run' nodes: 'Job jsh01d1q7f2c9', 'Job jsh01d1q7f2c5', and 'Job jsh01d1q7f2c4'. The third column has three 'Dataset' nodes: 'Inventory US East', 'Inventory US West', and 'Inventory US Central'. Arrows point from the first column to the second, and from the second to the third. A fourth 'Job / Job run' node, 'Job Inventory ETL', is positioned to the right of the third column, with arrows pointing to it from the 'Inventory US East', 'Inventory US West', and 'Inventory US Central' nodes. A right-hand panel is open, showing 'Inventory' details. It has tabs for 'LINEAGE INFO', 'SCHEMA', and 'HISTORY'. Under 'LINEAGE INFO', it lists 'TYPE: Glue table', 'LINEAGE NODE ID: 4qht0pp1q7f2cn', 'LINEAGE CREATED ON: Jul 14, 2023, 06:49:26 PM', and 'SOURCE ID: 4qht0pp1q7f2cn'. Below this, it shows 'METADATA FORMS (2)'. The first form is 'DatasetFormType' with 'DATASET TYPE: ASSET' and 'DATASET VERSION: latest'. The second form is 'Asset lineage form' with 'OWNING PROJECT ID: atsm4cne7297af', 'ASSET ID: 4bupowgjr1jfr', and 'ASSET TYPE: Glue table'.



Julia
Data engineer

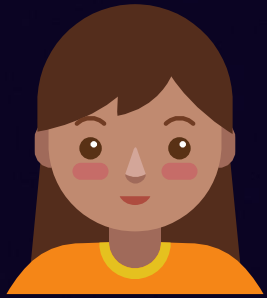
- Understand downstream dependencies
- See usage information of the data asset (i.e., used by 3 reports, 1 month ago)

Impact analysis

Julia starts with the data asset she would like to modify. From that asset she views downstream dependencies to understand impact.

The screenshot displays the AWS Glue console interface for the 'Inventory' data asset. The main view shows the lineage of the asset, starting from the 'Inventory' table (Technical name: inventory, Asset type: Glue Table) and branching into three subscription targets: 'Inventory_Web_Marketing_env', 'Inventory_Email_Campaign_env', and 'Inventory_Social_Campaign_env'. Each target has a 'Grant status' of 'Granted'. The 'Inventory_Web_Marketing_env' target is further linked to the 'Inventory_Web_Marketing_Model' dataset. A sidebar on the right provides details for the 'Inventory_Email_Campaign_env' subscription, including its type, grant status, and subscription target name.

SUBSCRIPTION INFO	
TYPE	Subscription target node
LINEAGE NODE ID	Subs0pp1q7f22q
GRANT STATUS	Subscribed Project
GRANT STATUS	Granted
SUBSCRIPTION TARGET NAME	inventory
GRANT ID	83jpowgjrj1jfr
ENVIRONMENT ID	envidwgjrj1jfr
EVENT TIMESTAMP	-
SOURCE ID	Subs0pp1q7f22q
SUBSCRIPTION DETAILS	
ASSET NAME	published-asset
ASSET ID	listowgjrj1jfr
OWNING PROJECT	Dataset Project
ASSET DESCRIPTION	This is a published asset description.



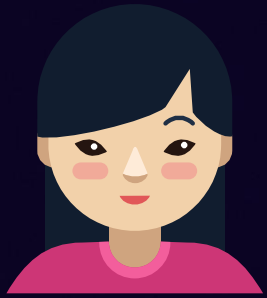
Julia
Data engineer

- Understand the report's upstream data
- Dive into what jobs were run and their status
- Look into the query that was executed to derive that data for the report

Troubleshooting analysis

Maria contacted Julia to report a data issue. Julia starts to investigate the data movement to understand what would have caused the issue.

The screenshot displays the AWS Glue console interface for the 'Inventory' table. At the top, there's a navigation bar with 'ADMIN', a search bar, and user information 'frontend-user'. Below this, a secondary navigation bar shows 'Admin', 'OVERVIEW', 'DATA', 'ENVIRONMENTS', and 'MEMBERS'. The main content area is titled 'Inventory' and includes technical details like 'Technical name: inventory' and 'Asset type: Glue Table'. A 'RE-PUBLISH ASSET' button is visible. The 'LINEAGE' tab is selected, showing a flow diagram. On the left, three 'Dataset' boxes represent 'Inventory US East', 'Inventory US West', and 'Inventory US Central', each with an event timestamp of Jul 12, 2023, 05:55:26 PM and 1 Column. These feed into a 'Job / Job run' box for 'Job Inventory ETL' with a status of 'COMPLETED' and an event timestamp of Jul 12, 2023, 05:58:26 PM. The job outputs to a 'Table / Dataset' box for 'Inventory' with an event timestamp of Jul 12, 2023, 05:55:26 PM and 15 Columns. This table is then published to a 'Table / AWS Glue / Published Inventory' with an event timestamp of Jul 14, 2023, 05:49:26 PM and 15 Columns. A 'View different versions to compare lineage' prompt is highlighted in orange. The footer contains 'Privacy | Site terms | Cookie preferences | © 2024 Amazon Web Services, Inc. or its affiliates. All rights reserved.' and a 'Send feedback' link.



Susan
Administrator

Regulatory compliance

Susan starts at the report detail page and selects the report column in question. She looks upstream to see how the column was calculated and from which sources to respond to auditors' queries.

- View lineage of data assets along with columns
- Traverse the lineage graph to view the upstream/downstream transformations for a column
- View snapshots of an asset to view how columns have changed over time

Inventory
Technical name: inventory - Asset type: Glue Table
This is the data asset for the Inventory table.

BUSINESS METADATA SCHEMA ASSET FILTERS DATA QUALITY LINEAGE Preview SUBSCRIPTION REQUESTS HISTORY

Inventory_Jul 14, 2023, 05:49:26 PM

Job / Job run
Job job6801q7f2c9
Status: COMPLETED
Event timestamp: Jul 12, 2023, 05:58:26 PM

Job / Job run
Job job6801q7f2c5
Status: COMPLETED
Event timestamp: Jul 12, 2023, 05:58:26 PM

Job / Job run
Job job6801q7f2c4
Status: COMPLETED
Event timestamp: Jul 12, 2023, 05:58:26 PM

Dataset
Inventory US East
Event timestamp: Jul 12, 2023, 05:55:26 PM
1 Column
Costs
boolean

Dataset
Inventory US West
Event timestamp: Jul 12, 2023, 05:55:26 PM
1 Column
Revenue
boolean

Dataset
Inventory US Central
Event timestamp: Jul 12, 2023, 05:55:26 PM
15 Columns
Costs
boolean
Revenue
string
Finance
boolean
Year
string
Month
string
Day
string
Name
string
Title
string
Position
string
Cash Flow
string

Job / Job run
Job Inventory ETL
Status: COMPLETED
Event timestamp: Jul 12, 2023, 05:58:26 PM

Base
Inventory
Event timestamp: Jul 14, 2023, 05:49:26 PM
15 Columns
Costs
boolean
Revenue
string
Finance
boolean
Year
string
Month
string
Day
string
Name
string
Title
string
Position
string
Cash Flow
string

Table / AWS Glue / Published
Inventory
Event timestamp: Jul 16, 2023, 05:49:26 PM
13 Columns
Costs
boolean
Revenue
string
Finance
boolean
Year
string
Month
string
Day
string
Name
string
Title
string
Position
string
Cash Flow
string

Subscription target / AWS Glue
Inventory_Web_Marketing_env
Grant status: Granted

Subscription target / Redshift cluster
Inventory_Email_Campaign_env
Grant status: Granted


Subscription target / Redshift cluster
Inventory_Social_Campaign_env
Grant status: Granted

Privacy | Site terms | Cookie preferences | © 2024 Amazon Web Services, Inc. or its affiliates. All rights reserved. Send feedback

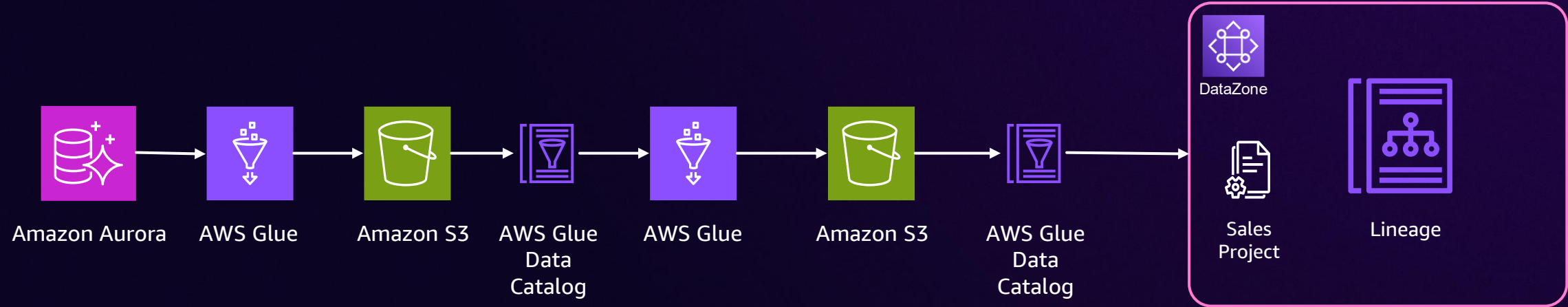
Data lineage demo



Demo 1 – AWS Glue ETL

 Central governance account


Data Producer
(Sales)





SALES

Search data

ASSET PUBLISHED
Total Quantity Per Buyer
Technical name: total_quantity_per_buyer Asset type: Glue Table
No description

ACTIONS

BUSINESS METADATA SCHEMA ASSET FILTERS DATA QUALITY LINEAGE SUBSCRIPTION REQUESTS HISTORY

SUMMARY EDIT

The "total_quantity_per_buyer" table is a crucial data source for retail businesses, providing valuable insights into the purchasing behavior and demographics of their customers. This table contains information about the total quantity of products purchased by each customer or user, along with their associated user ID, username, city, and state.

The city and state columns provide important geographic information about the customer's location, which can be leveraged by retail managers and marketers to understand regional trends, target marketing campaigns, and optimize product distribution. The total_quantity column represents the cumulative quantity of products purchased by each customer across all their transactions, offering a comprehensive view of their overall purchasing activity.

The userid column serves as a unique identifier for each customer or user, enabling the tracking of individual purchasing patterns and the ability to personalize the customer experience. The username column provides an additional layer of customer information, potentially allowing for more targeted marketing efforts or the identification of influential brand advocates.

This table can be instrumental in various retail-related analyses and decision-making processes. Retail managers can use the data to identify high-volume customers, understand regional purchasing preferences, and allocate resources more effectively. Marketers can leverage the insights to develop targeted campaigns, personalize product recommendations, and enhance customer engagement. Analysts can utilize the data to uncover trends, forecast demand, and make informed strategic decisions that drive business growth and profitability.

Use Case

The "total_quantity_per_buyer" table can be utilized by various retail professionals and organizations to drive strategic decision-making and enhance overall business performance. This table can be particularly valuable for:

- Product Listing and Inventory Management:** Retail managers and product planners can leverage the data to identify high-volume customers and their purchasing preferences, allowing them to optimize product assortment, pricing, and inventory levels to meet customer demand more effectively.
- Sales Analysis and Forecasting:** Retail analysts and business intelligence teams can use the data to uncover regional sales trends, forecast future demand, and make informed decisions about resource allocation, marketing strategies, and expansion plans.
- Customer Engagement and Loyalty Programs:** Retail marketers and customer experience professionals can utilize the data to personalize product recommendations, develop targeted marketing campaigns, and enhance customer loyalty by catering to the unique needs and preferences of high-volume buyers.

The data in this table can benefit a wide range of retail-related organizations, including brick-and-mortar stores, e-commerce platforms, and omnichannel retailers. It can also be valuable for industry analysts, consultants, and researchers studying consumer behavior and trends in the retail sector.

[SHOW LESS](#)

README EDIT

How to Use Access the Data: Use your organization's data platform to locate and access the Total Quantity Per Buyer asset. Run Queries: Execute SQL or equivalent queries to retrieve aggregated data. Apply Filters: Customize the dataset by applying relevant filters (e.g., product type or time frame). Export Results: Save the queried data in your preferred format for downstream applications.

[SHOW LESS](#)

GLOSSARY TERMS [What's this?](#)

No terms found
Annotate this asset with terms

[ADD TERMS](#)

ASSET DETAILS

DATA QUALITY SCORE

100

[View data quality details](#)

PUBLISHED REVISION

7

CATALOG VISIBILITY

Searchable EDIT

LATEST INVENTORY REVISION

7

OWNING PROJECT

sales

DOMAIN UNIT

corporate

SUBSCRIPTION APPROVAL

Required EDIT

LAST MODIFIED BY

WSParticipantRole

UPDATED AT

Nov 25, 2024, 03:10:30 PM

CREATED BY

SYSTEM

CREATED AT

Nov 25, 2024, 03:01:47 PM

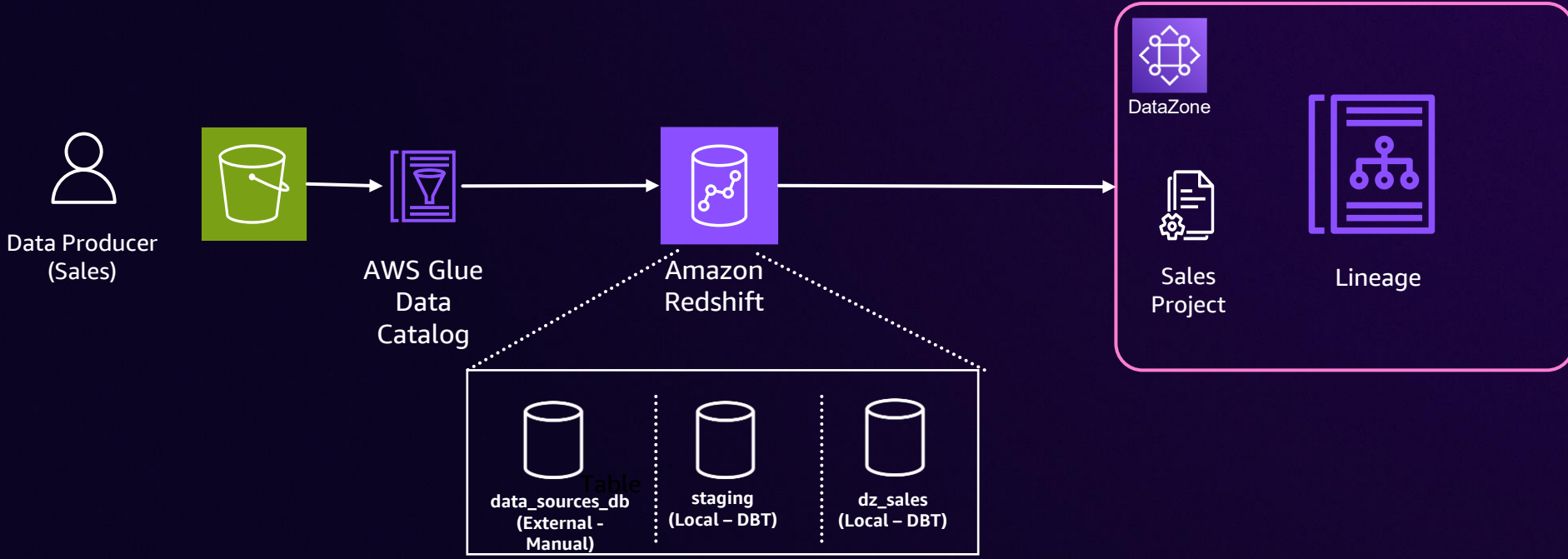
SUBSCRIBER COUNT

1

Demo 2 – Integration with DBT



Central governance account



← Back

ASSET PUBLISHED

Sales By Date

Technical name: sales_by_date • Asset type: Redshift Table

No description

ACTIONS ▾

BUSINESS METADATA SCHEMA ASSET FILTERS DATA QUALITY LINEAGE SUBSCRIPTION REQUESTS HISTORY

SUMMARY

EDIT

The 'sales_by_date' table is a comprehensive dataset that captures the daily sales performance of a retail business. This table provides a detailed record of the quantity sold, revenue generated, and various temporal attributes associated with each sales transaction.

The quantity_sold column represents the total number of units of a product or products sold on a given date, offering insights into customer demand and purchasing patterns. The revenue column, on the other hand, reflects the total monetary value of the sales on a particular day, enabling analysis of the financial performance and profitability of the business.

The caldate column provides the calendar date associated with the sales data, typically in the format YYYY-MM-DD. This information is crucial for understanding the temporal dynamics of sales, such as seasonal trends, holiday impacts, and day-of-the-week effects.

The week, month, year, and qtr columns further categorize the sales data, allowing for analysis of sales performance across different time frames. This granular temporal information can be valuable for retail managers and analysts in identifying patterns, forecasting demand, and aligning marketing and operational strategies.


The day column indicates the day of the week (e.g., Monday, Tuesday) corresponding to the sales data, which can be useful for understanding the impact of weekday versus weekend sales. The holiday column serves as a flag, signaling whether the sales date coincides with a recognized holiday (e.g., Christmas, Thanksgiving), enabling the analysis of the influence of special events on sales.

Overall, the 'sales_by_date' table provides a comprehensive view of a retail business's sales performance, offering valuable insights that can inform strategic decision-making, optimize inventory management, and enhance marketing efforts to drive growth and profitability.

SHOW LESS ▾

README

HIDE ▾

 **Add a readme section**
A readme will help users understand this page better

[CREATE README](#)

ASSET DETAILS

DATA QUALITY SCORE



[View data quality details](#)

PUBLISHED REVISION

6

CATALOG VISIBILITY

Searchable

EDIT

LATEST INVENTORY REVISION

6

OWNING PROJECT

sales

DOMAIN UNIT

corporate

SUBSCRIPTION APPROVAL

Required

EDIT

LAST MODIFIED BY

SYSTEM

UPDATED AT

Nov 25, 2024, 04:05:14 PM

CREATED BY

SYSTEM

CREATED AT

Nov 25, 2024, 04:02:11 PM

SUBSCRIBER COUNT

0






GLOSSARY TERMS [What's this?](#)

SDG&E's data governance journey



About San Diego Gas & Electric

San Diego Gas & Electric is an innovative San Diego-based energy company that provides clean, safe, and reliable energy to better the lives of the people we serve in San Diego and southern Orange counties

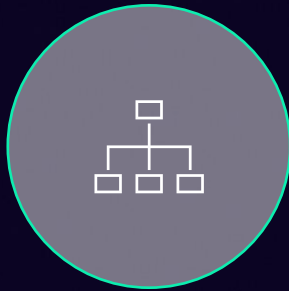
 4,000+ employees	 3.7M customers	 1.5M electric meters	 905K gas meters	 17.4K power line miles
---	---	---	--	---

 4,100 sq. mi. in San Diego and S. Orange counties
--

 44.5% energy from renewable sources
--



Data governance journey at SDG&E



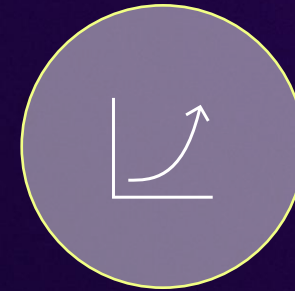
Data mesh architecture

- Deployed a data mesh architecture on AWS
- Optimized data pipelines to enable efficient data production and consumption, starting with asset management and customer service



Governance and operating model

- Enhanced data governance via hub and spoke model
- Aligned centralized IT to enable consistency and best practices and business-aligned IT to drive decentralized agility

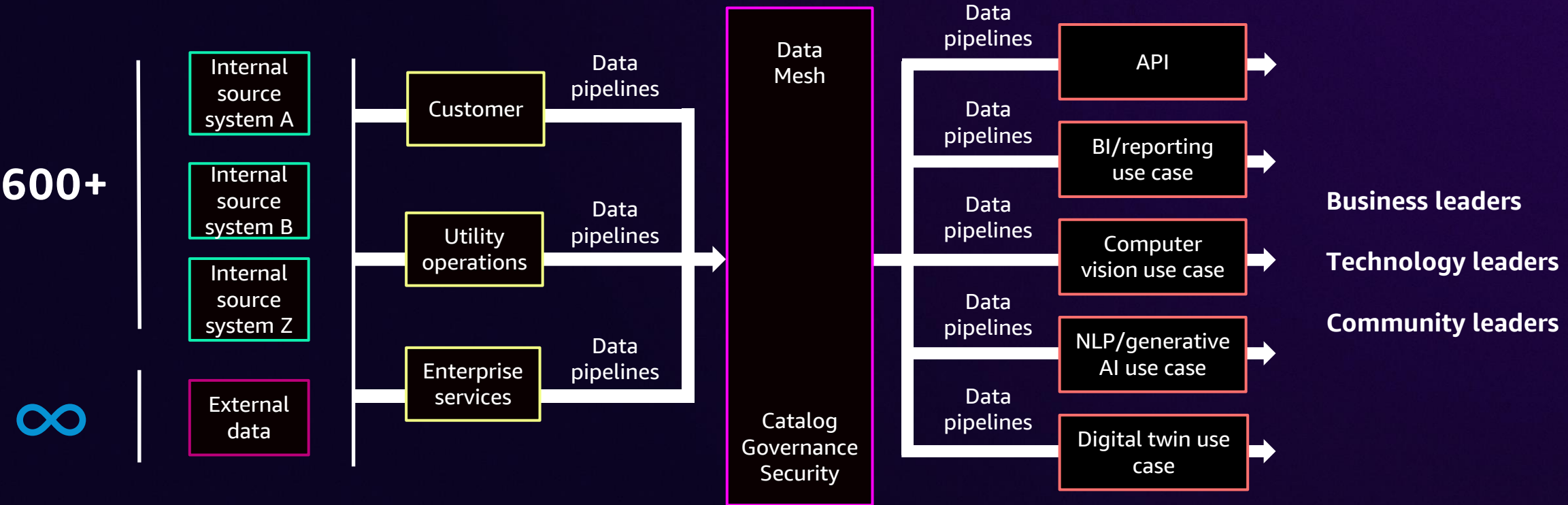


Innovation acceleration

- Accelerated innovation velocity, emphasizing data and IaC reusability within the data mesh and architecture
- Building new capabilities using serverless, cloud native services to quickly respond to changing business conditions

Governing data through the mesh

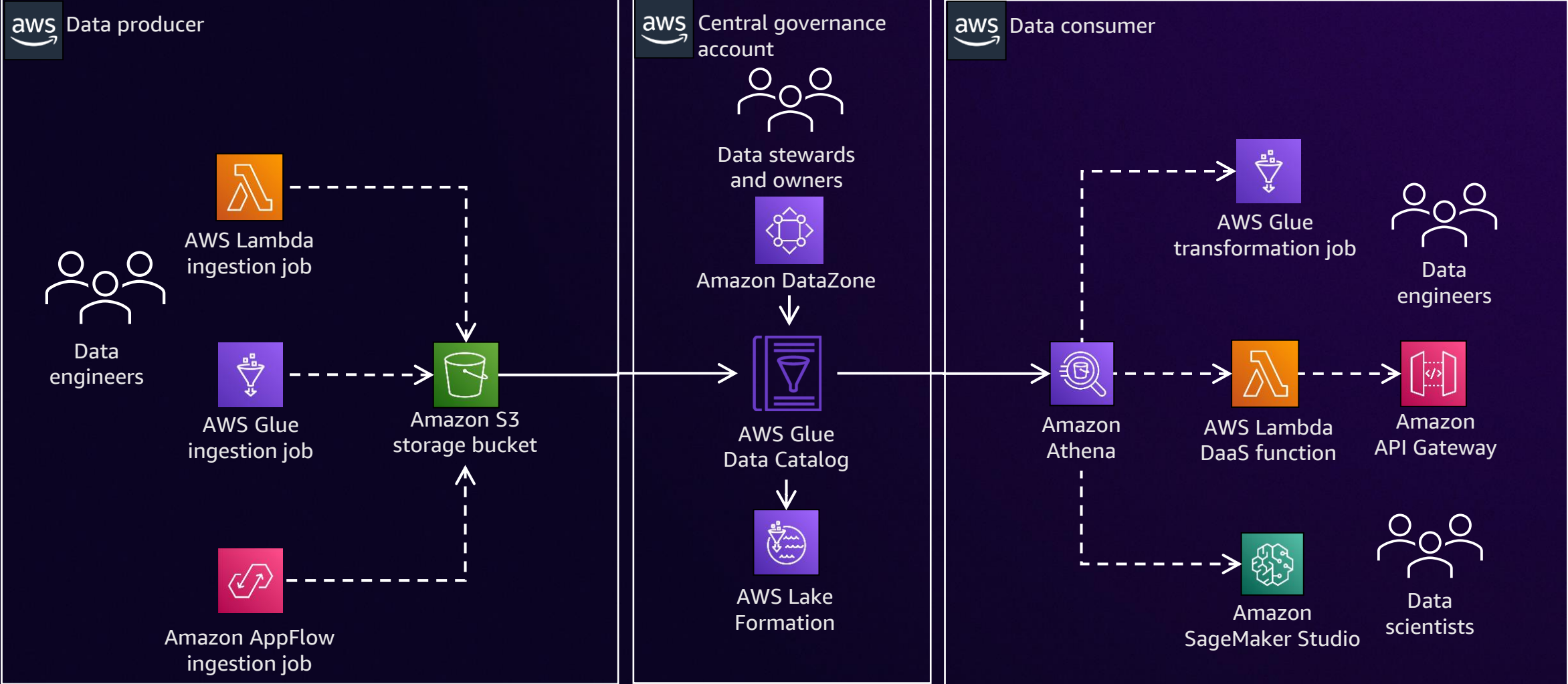
Source system data → Data producers → Data products ↔ Data consumers → Insights/decisions



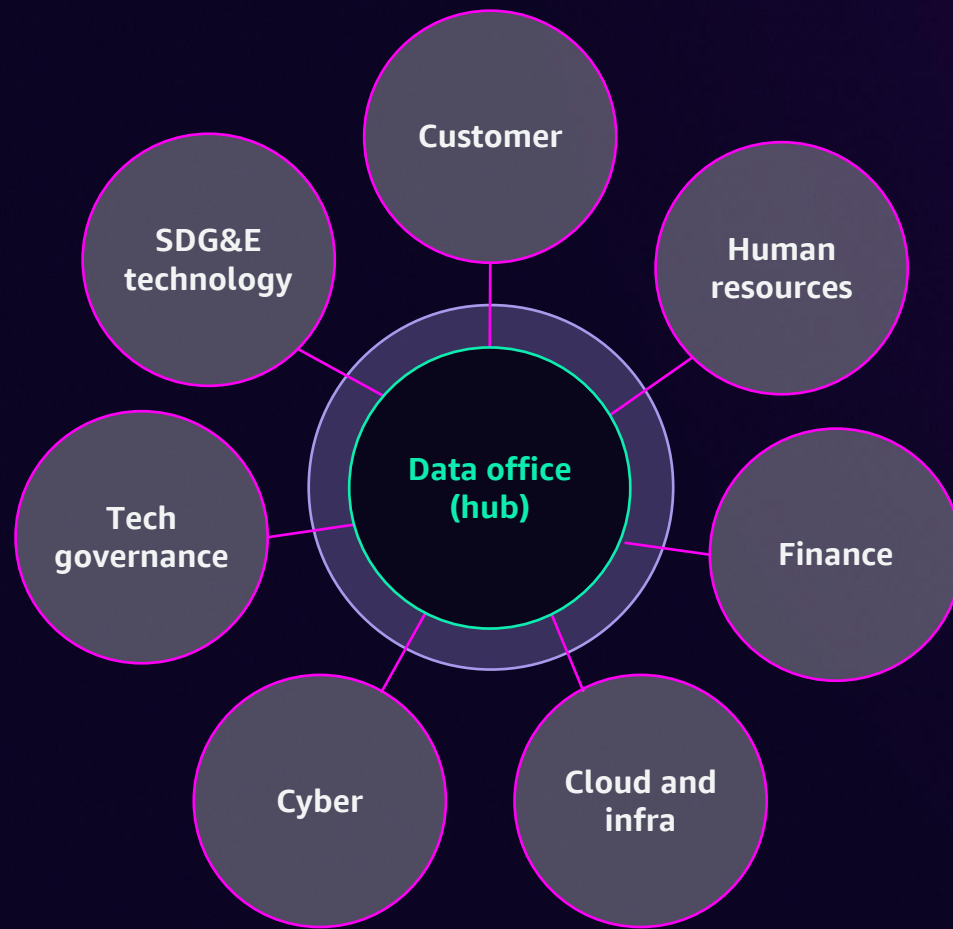
Integrate once, consume indefinitely



Our solution architecture



How SDG&E organizes for data governance

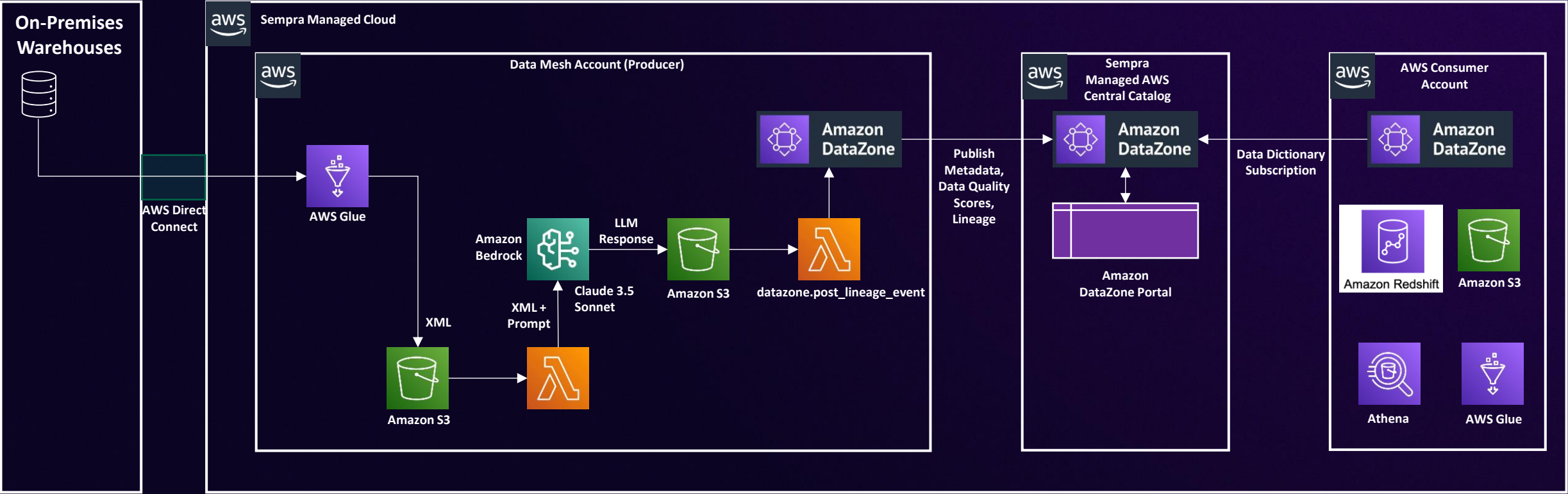


Hub is a centralized organization led out of the Chief Information Officer/Chief Digital Officer organization to provide standardized data and AI services across the enterprise

Spokes are dedicated data and AI groups for the businesses that drive domain-specific initiatives while using the services from the hub

Blue area includes activities beyond the purview of both the hub and the spokes, but successful delivery of data and AI projects vitally depend on them

Innovation: On-premises metadata lineage in Amazon DataZone

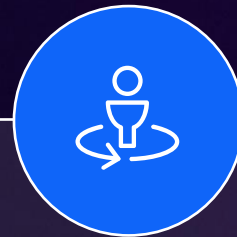


How SDG&E measures data governance success



Data mesh and data usability

- Data quality
- Data discoverability and accessibility
- Speed to onboard new data product
- Data mesh adoption (both producers and consumers)



People and culture

- Organizational adoption
- Standards and guidelines published
- Federated data teams delivering persistent value
- Data evolves into a base capability rather than just an asset



Accelerating our speed to innovate

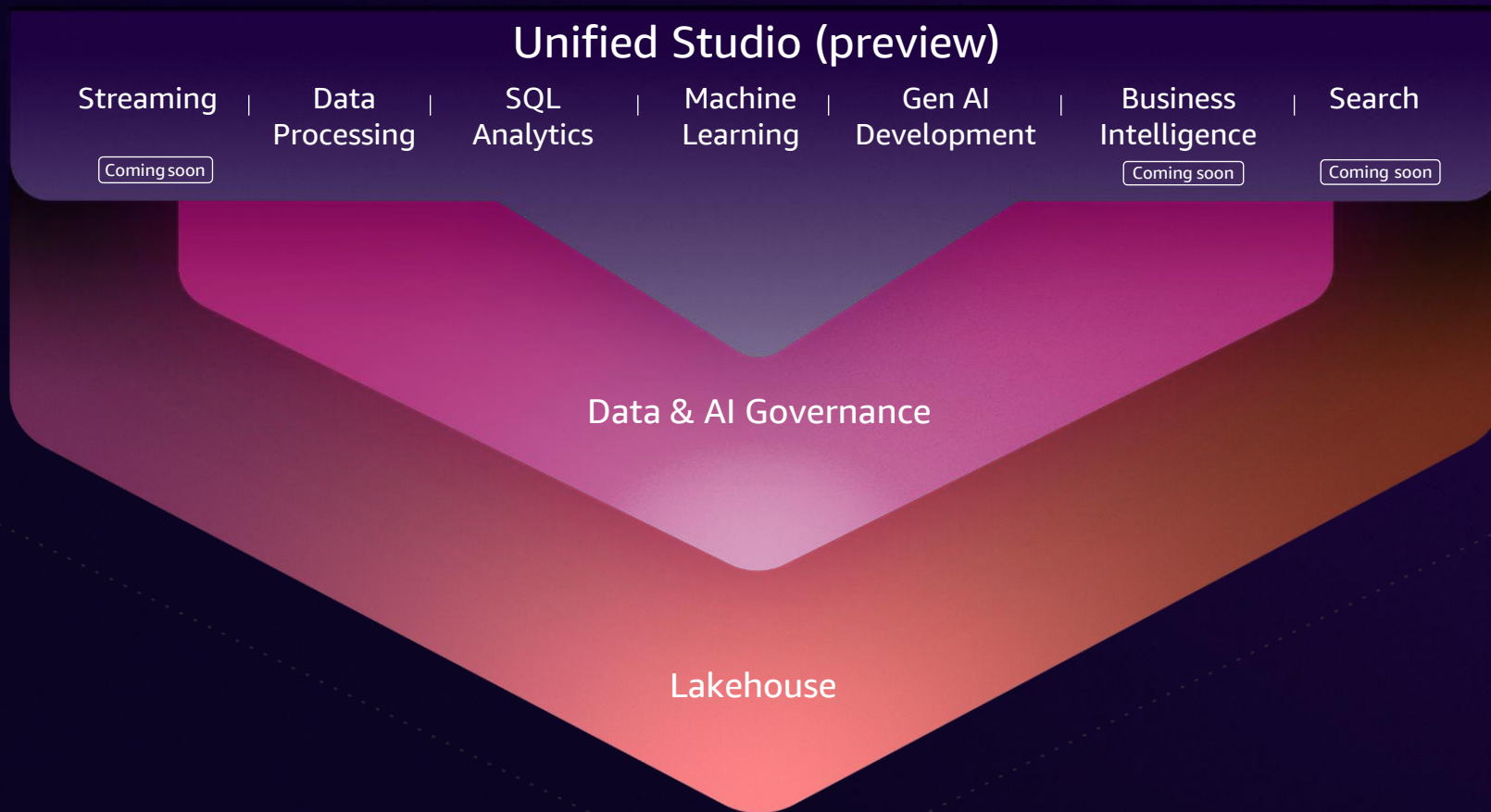
100%
increase

Data lineage in Amazon SageMaker Unified Studio



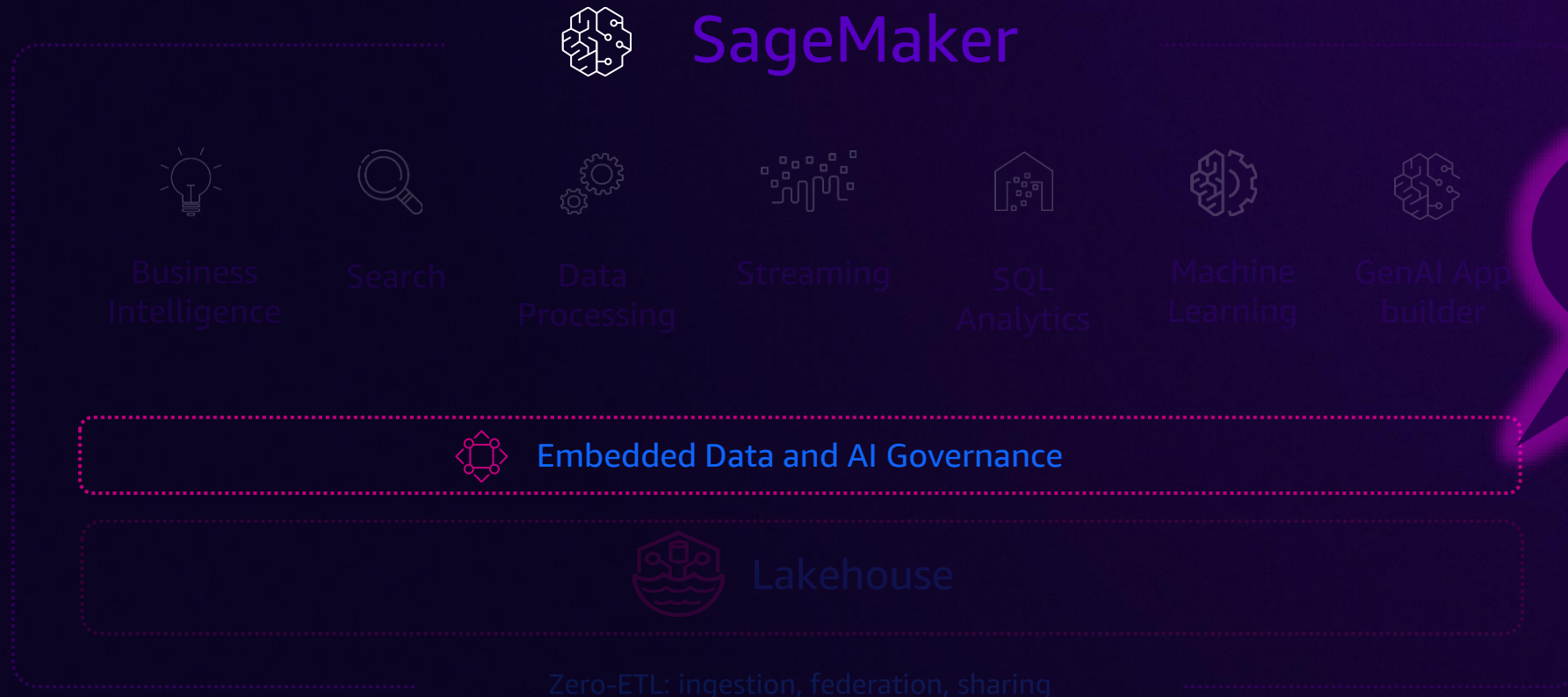
Your center for data, analytics & AI

Amazon SageMaker



Amazon SageMaker

SINGLE DEVELOPMENT ENVIRONMENT TO USE ALL YOUR DATA AND TOOLS FOR ANALYTICS AND AI



Built on
Amazon
DataZone

Amazon SageMaker Unified Studio

DISCOVER, GOVERN, AND COLLABORATE ON DATA AND AI SECURELY, WITH A UNIFIED CATALOG

Sales_2024
Technical name: [Technical Name] • Domain unit: [Domain unit] [---] • Asset type: Glue Table
Description of all the information and artifacts required for creation of Sales dataset for the organization.

BUSINESS METADATA SCHEMA DATA QUALITY LINEAGE **Preview**

Search displayed nodes

Product_info
Technical name: [Technical Name]
Event timestamp: Jan 21, 2024, 10:10:12 AM
3 Columns

Customer_sales
Technical name: [Technical Name]
Event timestamp: Jan 21, 2024, 10:10:12 AM
8 Columns

Sales_2024
Technical name: [Technical Name]
Event timestamp: Jan 22, 2024, 12:10:12 AM
6 Columns

Product_info
Technical name: [Technical Name]
Event timestamp: Jan 22, 2022, 12:12:12 AM
3 Columns

Customer_sales
Technical name: [Technical Name]
Event timestamp: Jan 22, 2022, 12:12:12 AM
8 Columns

Sales_2024
Technical name: [Technical Name]
Event timestamp: Jan 22, 2022, 12:12:12 AM
6 Columns

Base
AWS Glue / Table
Sales_2024
Technical name: [Technical Name]
Latest data quality score: 88.89
Event timestamp: Jan 22, 2022, 12:12:12 AM
10 Columns

Data Lineage

Built on Amazon DataZone



Resources



OpenLineage.io



Amazon DataZone
data lineage blog

Thank you!

Priya Tiruthani
tirutn@amazon.com

Harel Shein
OpenLineage TSC

Rob Malowney
SDG&E

Leo Gomez
golonar@amazon.com



Please complete the session survey in the mobile app