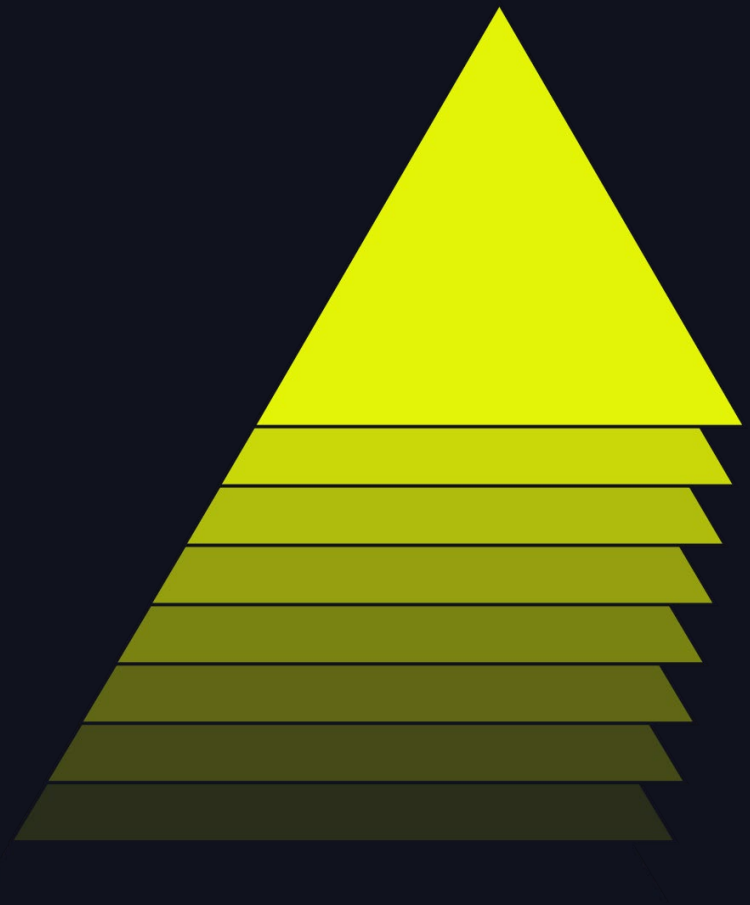
DATA+AI SUMMIT
BY databricks

Democratize Data Discovery And Data Insight with Databricks



Tao Feng, Aly Hirani
06/2024

Product safe harbor statement

This information is provided to outline Databricks' general product direction and is for **informational purposes only**. Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks discretion and may not be delivered as planned or at all

About Me

Tao Feng

- Senior Staff TLM at Databricks
- Working on Data Discovery and Lineage
- Co-Creator of Amundsen (3.5k+ github star) and Apache Airflow PMC
- Previously worked at Lyft, and various other tech companies

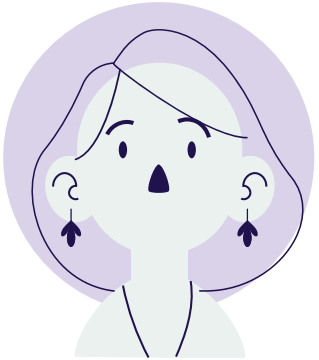
Aly Hirani

Data “Trust” Discovery

Data Personas

Data Producers / Data Stewards

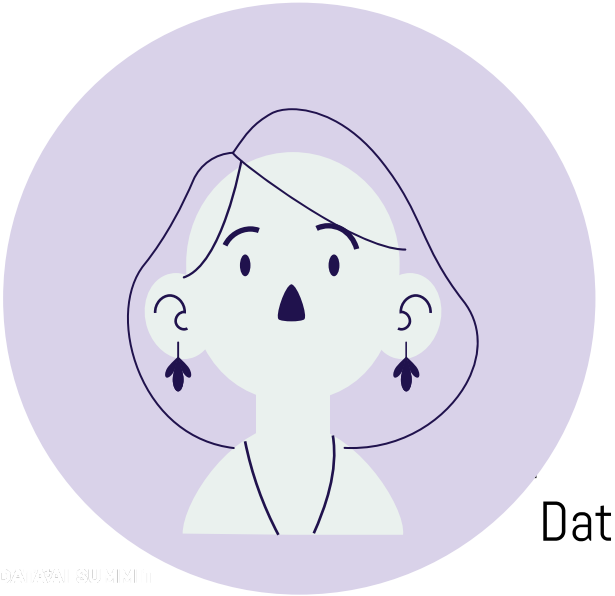
Create certified data, Publish certified data, Notify downstream users on data quality issues, Grant access, Answer questions around data



Data Consumers

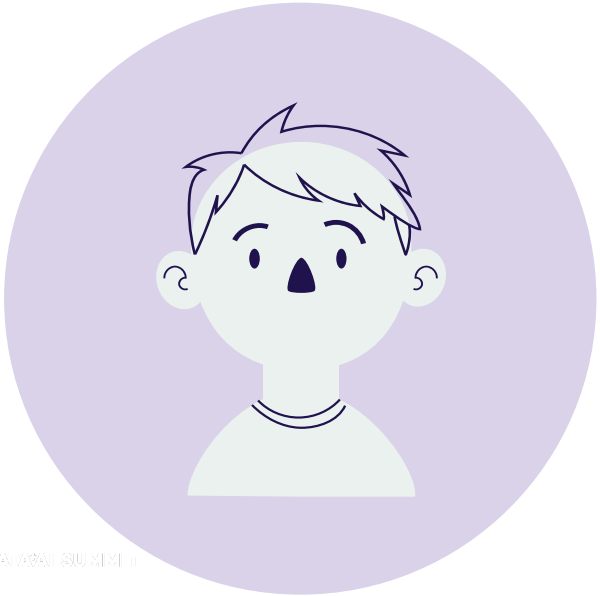
Explore data, Create biz dashboards / reports / notebooks or train models

*I've been looking for **trust** data for my analysis, but I don't know where they are or whether the data I found is trustworthy.*



Data Consumer / Business Analyst

*I've built a new certified dataset and would like to **migrate** existing users from legacy dataset to this new one.*



Data Steward

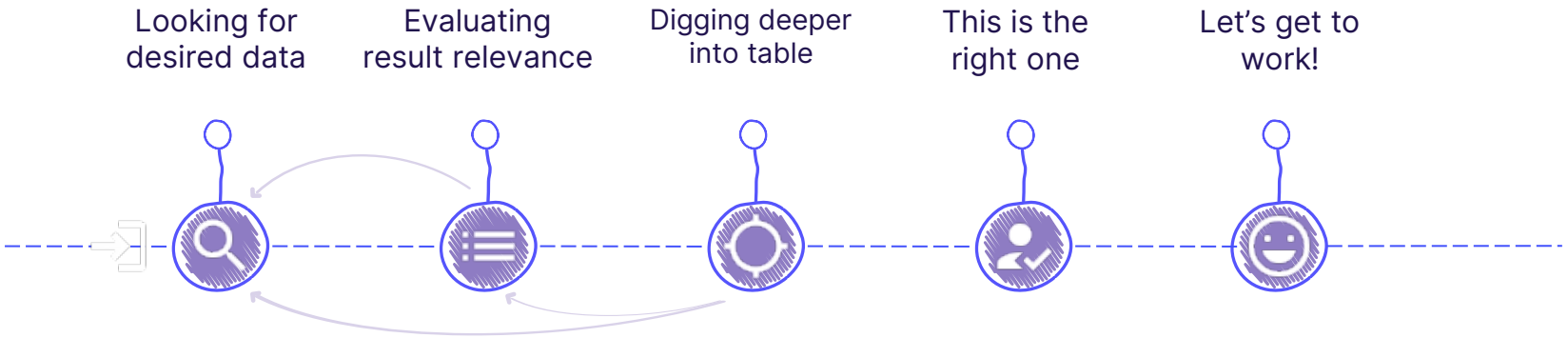
Challenge

What main challenges are we facing in discovery CUJs

There is a data trust gap in between the data producer and data consumer

- (*Data Consumer*) Poor understanding of the data leads to lower data quality, lower productivity, high risk of duplicate work and mistakes. This could lead to make wrong decisions for critical biz analysis.
- (*Data Producer*) Fail to notify critical dataset quality issues or migrate data users from legacy dataset to certified dataset make data trust even worse.

Data Consumer User Journey



Search And Browse

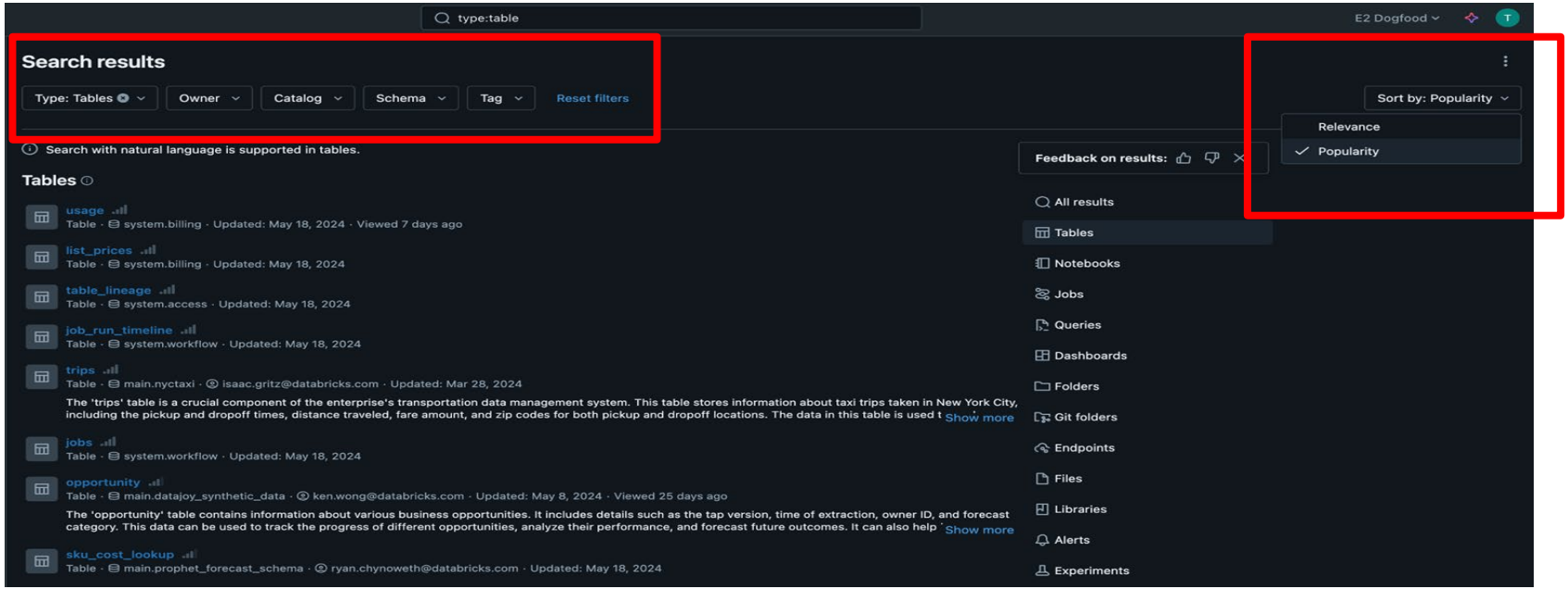
Search

- Discovery new data
  - Based on business context (descriptions and tags)
- Find known or used data
  - Based on source table name

Browse

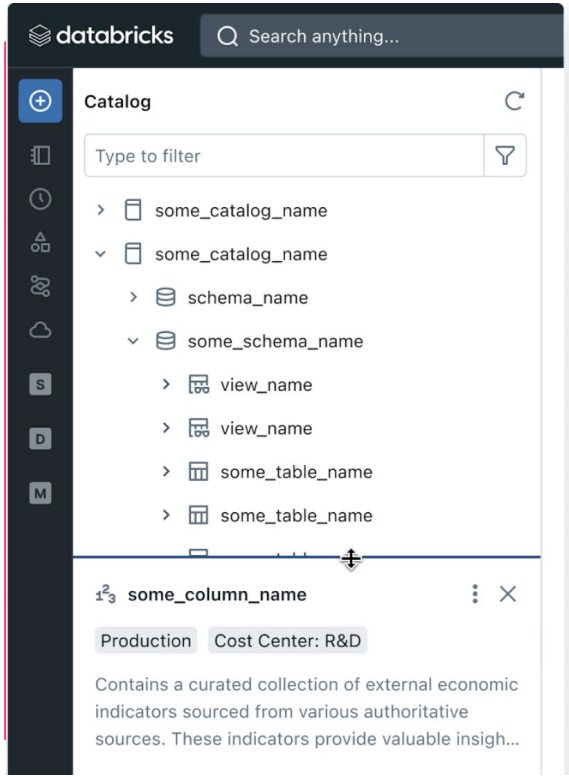
- Browse all recent/favorite/Popular tables

Global Search



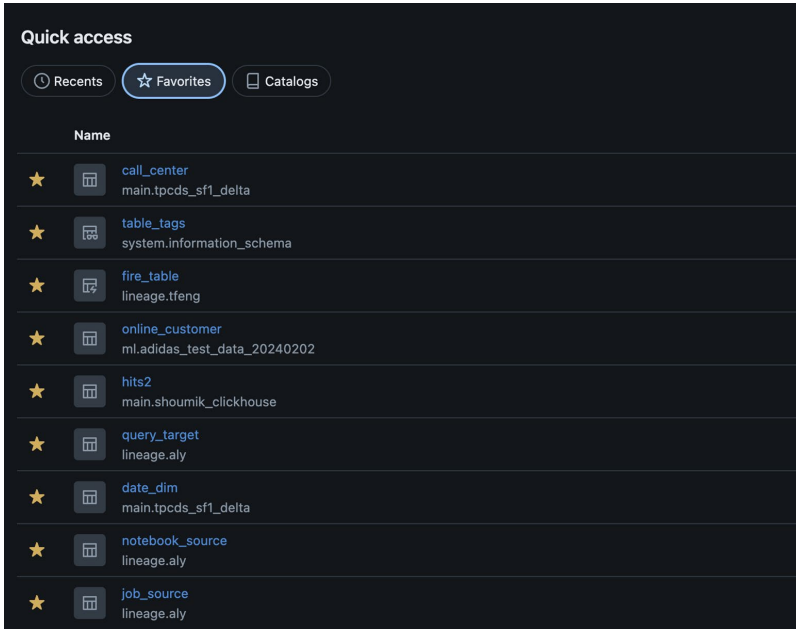
Schema Browser

- In-context data discovery without switch back to Catalog Explorer
- Surface active / favorite tables
- (In Future) Show metadata information for the selected table



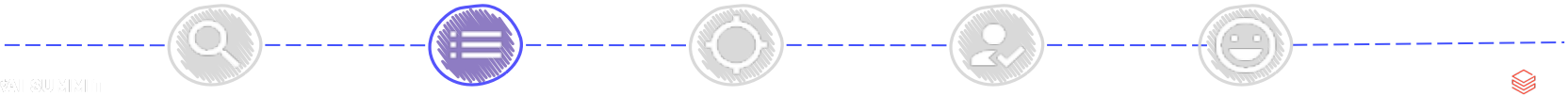
Catalog Explorer

- Recent/Favorite



Identify Most Relevant Results

- Global popularity: Accessed frequently overall
- Personal popularity: Accessed frequently by user, her team, or teams close to hers
- Personal relevance: Based on previous tables used, tables from the same source or having similar business context (e.g., via tags)
- # of bookmarks / favorites
- Official vs. user-generated tables
- Having quality rating, descriptions and other metadata, and complete date range

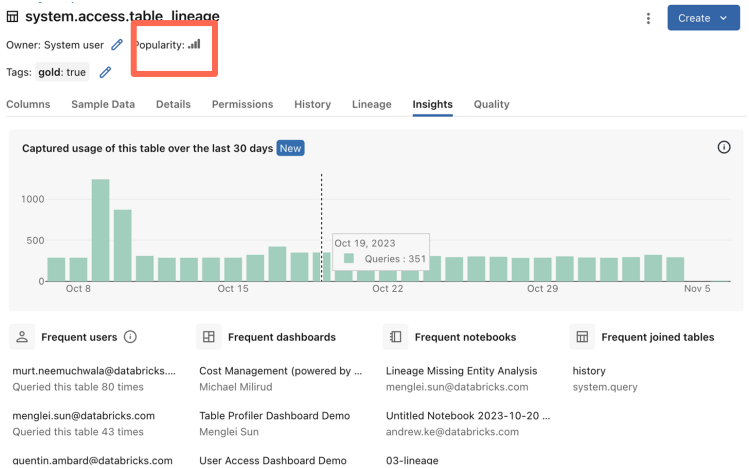


Explore data

CUJ: Evaluate data trustworthiness

Scenario: Rebecca wants to assess the trustworthiness of a dataset by reviewing its quality metrics, popularity usage, and lineage information before incorporating it into her data science project.

User Goals: Evaluate trust signals to ensure data trustworthy before used for analysis.



Trust Indicators

Official tables

Clear and complete description and tags

Well-documented changes (transformation model)

Having quality rating

Knowing the creator (team/person)

What's the source / downstream usage? Any external or manually updated data?

Frequently used

Recognising frequent users

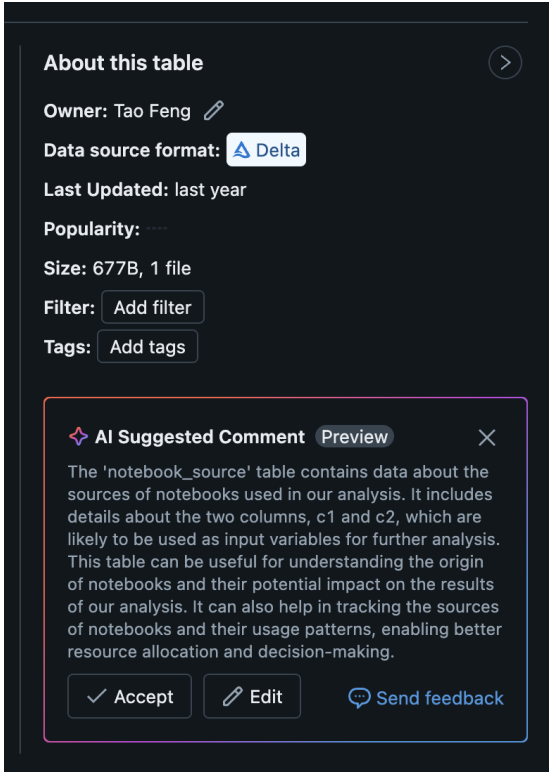
Updated recently





Explore Data

- Introduce overview tab which surfaces the trust indicators/signals
- The same trust information will be available in different discovery surface areas: search, authoring interface, assistant

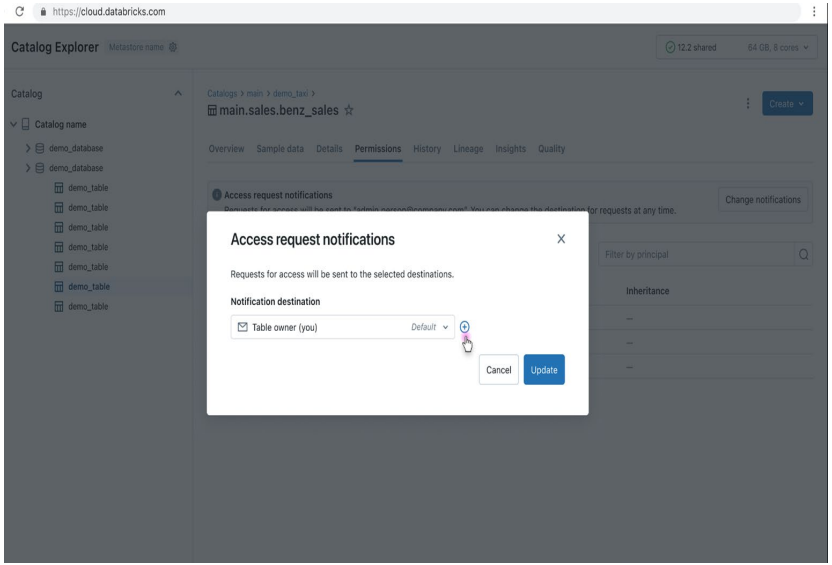


Request Access

CUJ: Request data access on a known table

Scenario: Alex only has BROWSE permission (no SELECT permission) and wants to request access a table after previous evaluation

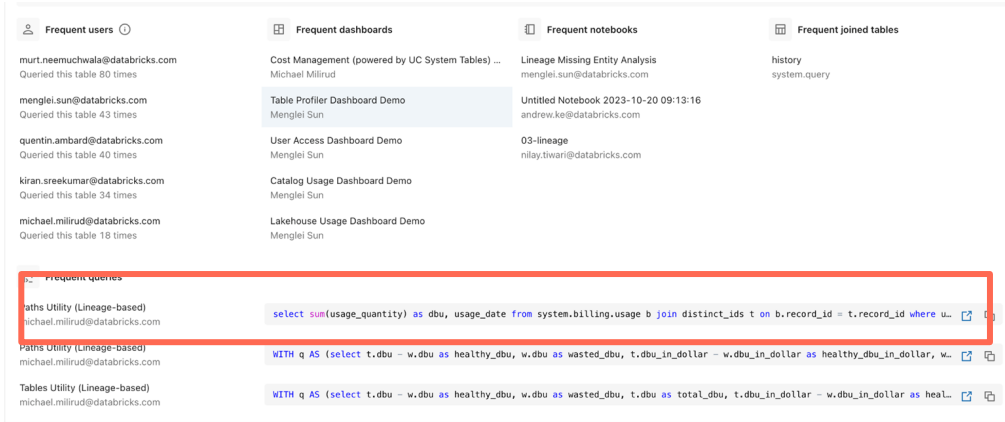
User Goals: Get select access on a known table

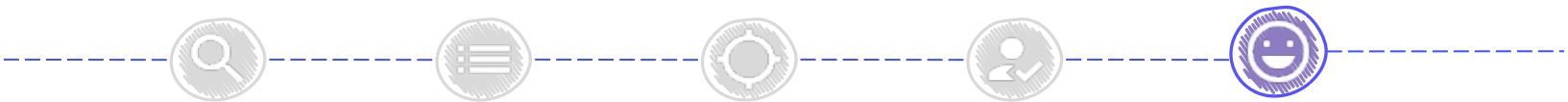


Use Data

CUJ: Use data to create query/notebook/dashboard based on known examples

Scenario: John wants to start creating query / notebook / dashboard with known data / example
User Goals: User produces data analysis with good data





End-to-End Lineage and Insight

Lineage Today

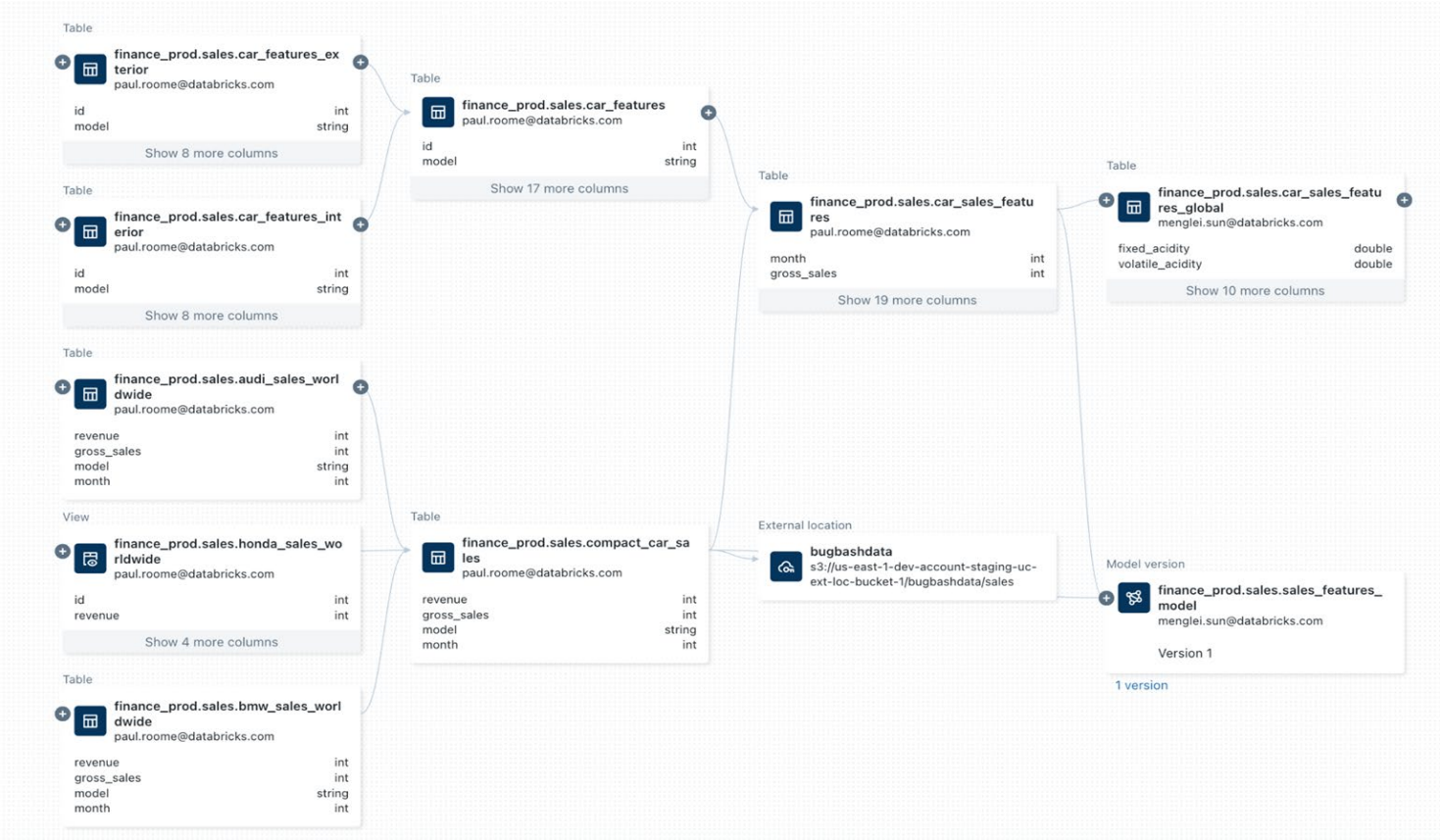
Automatic Capturing

- Unity Catalog
  - Tables
  - Paths
  - Volumes
  - Model Versions
  - Functions
- Notebooks
- Dashboards
- Queries
- Workflows
- Delta Live Tables

Questions Answered

- Can I trust this data?
- Why is this data missing/inaccurate?
- Who has seen this PII data?
- Where did this data come from?
- Can I deprecate this table/column?

Lineage Graph Today



Today, lineage has great coverage within Databricks

But what about everything else?

Bring Your Own Lineage

External Sources

Introducing…

Bring Your Own Lineage!

- Define custom ‘entities’ for any external sources
- Define custom ‘relationships’ between custom entities and Databricks assets
- API support to create, update, and delete custom entities and relationships
- View the new lineage in the UI

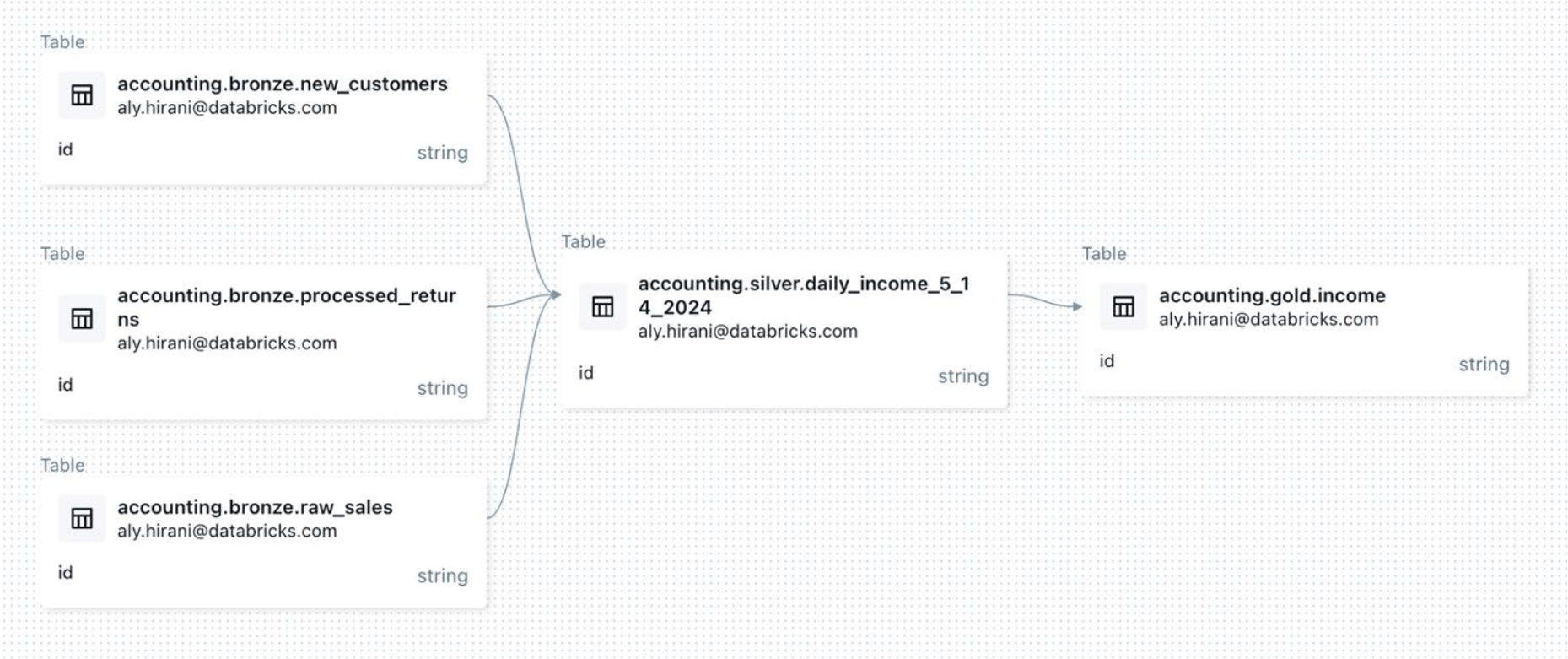
Why?

Tell the full story!

- Impact analysis
  - Standard BI tools
- End to end data provenance
  - ‘True’ origin
- Fix broken lineage
  - Staging tables

Demo !

Before…



Example

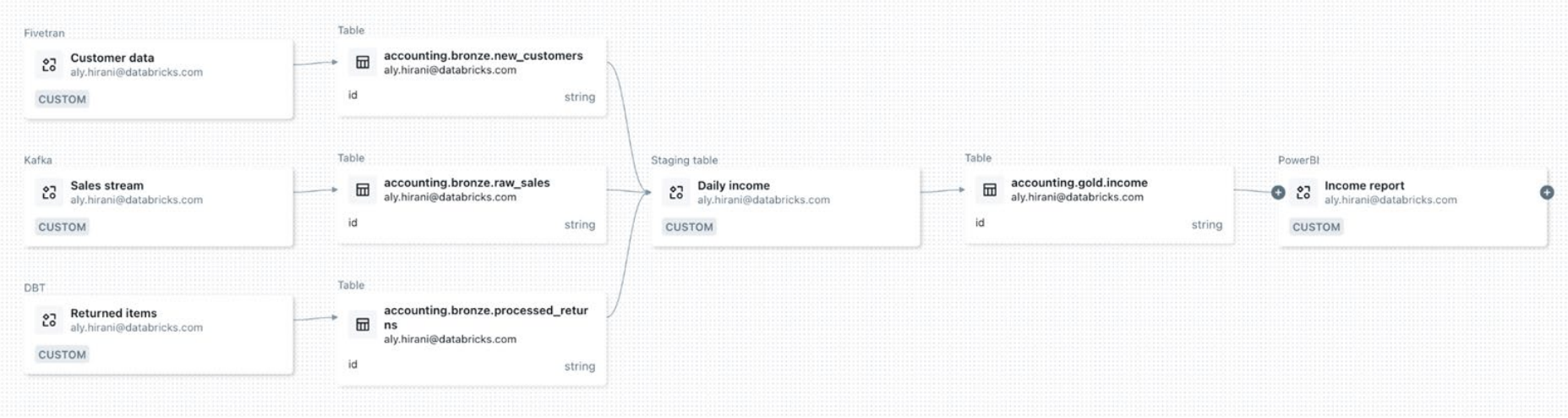
Entity

```
{
  "entity_id": {
    "provider_type": "CUSTOM",
    "guid": "kafka-customers"
  },
  "entity_type": "Kafka Stream",
  "display_name": "Customer Logs",
  "url": "https://www.kafka.com",
  "description": "Some important Kafka topic",
  "properties": "{"checkpoint": "/mnt/1234"}"
}
```

Relationship

```
{
  "source": {
    "provider_type": "CUSTOM",
    "guid": "kafka-customers"
  },
  "target": {
    "provider_type": "DATABRICKS",
    "databricks_type": "TABLE",
    "guid": "sales.bronze.customers_raw"
  }
}
```

After!



Next Steps

- Open source SDK
- Community driven integrations
  - Crawl lineage from external systems to ingest into Databricks
- Automatic capturing for first party ingestion