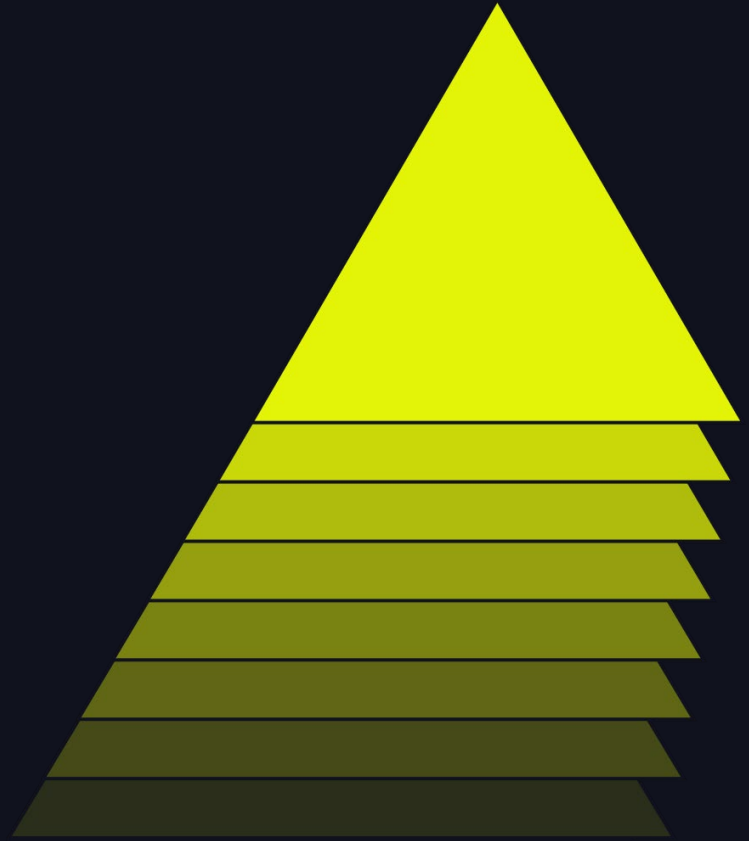


Democratize Data Discovery And Data Insight with Databricks

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



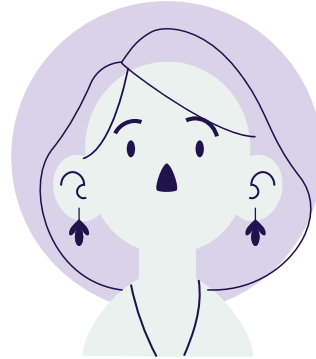
Data "Trust" Discovery

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



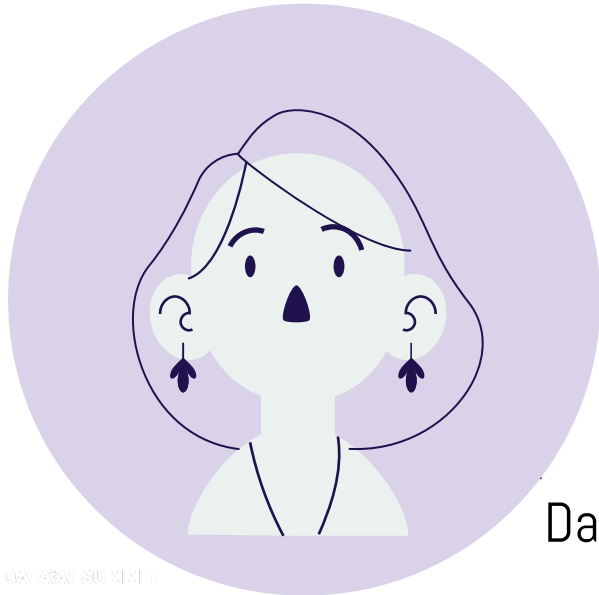
Data Consumers

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.*

”

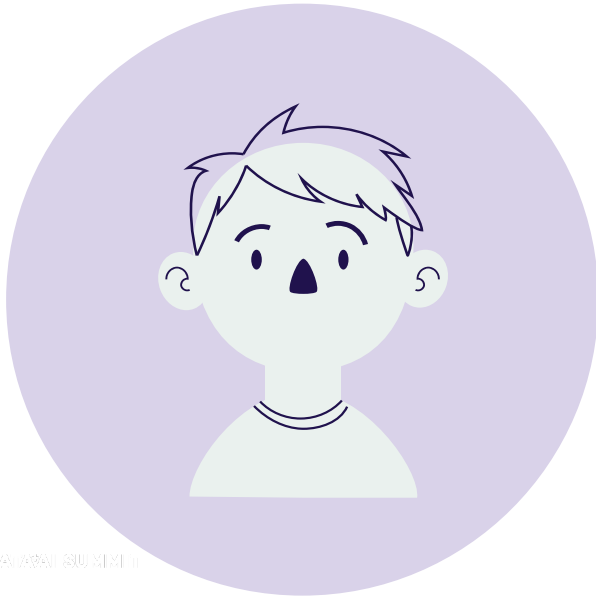


Data Consumer / Business Analyst

“

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

”



Data Steward

Challenge

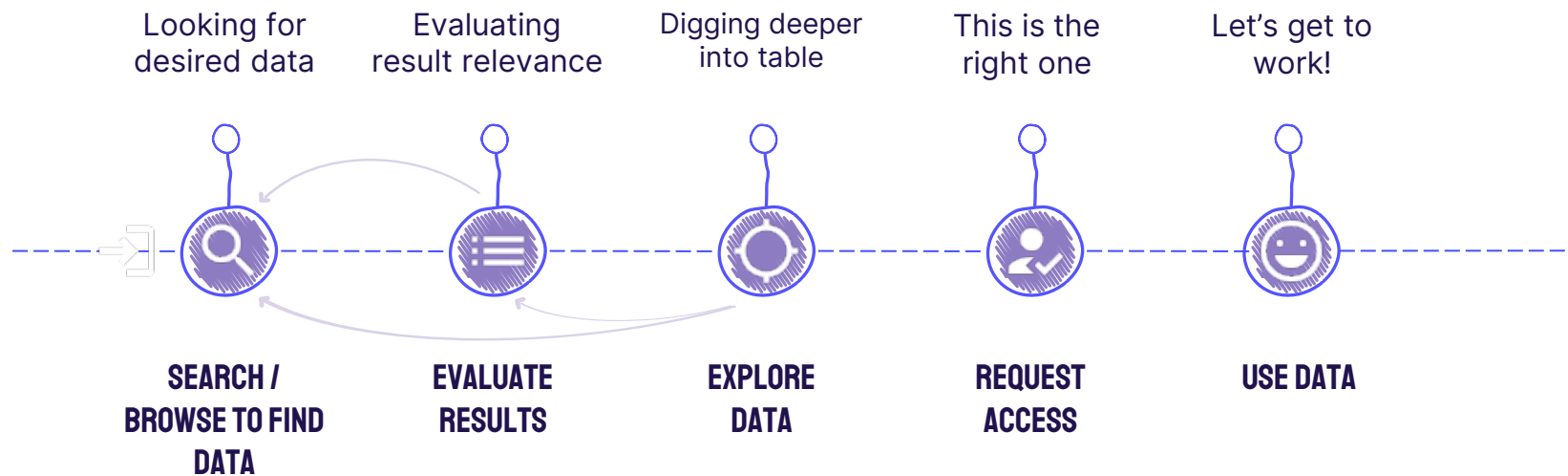
What main **challenges** are we facing in discovery CUJs



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

The screenshot displays the Databricks Global Search interface. At the top, a search bar contains the query "type:table". The user's name "E2 Dogfood" is visible in the top right corner. The search results are displayed in a list format under the heading "Search results".

Below the search results heading, there are several filter buttons: "Type: Tables", "Owner", "Catalog", "Schema", and "Tag", along with a "Reset filters" link. A red box highlights this filter area.

Below the filters, there is a message: "Search with natural language is supported in tables."

The search results are listed under the heading "Tables". The results include:

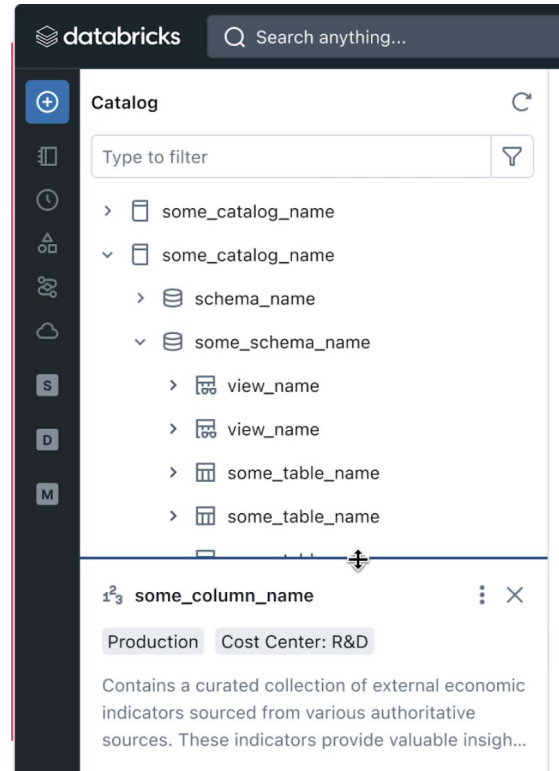
- usage** .tbl - Table - system.billing - Updated: May 18, 2024 - Viewed 7 days ago
- list_prices** .tbl - Table - system.billing - Updated: May 18, 2024
- table_lineage** .tbl - Table - system.access - Updated: May 18, 2024
- job_run_timeline** .tbl - Table - system.workflow - Updated: May 18, 2024
- trips** .tbl - Table - main.nyctaxi - Isaac.gritz@databricks.com - Updated: Mar 28, 2024
The 'trips' table is a crucial component of the enterprise's transportation data management system. This table stores information about taxi trips taken in New York City, including the pickup and dropoff times, distance traveled, fare amount, and zip codes for both pickup and dropoff locations. The data in this table is used to [Show more](#)
- jobs** .tbl - Table - system.workflow - Updated: May 18, 2024
- opportunity** .tbl - Table - main.datajoy_synthetic_data - ken.wong@databricks.com - Updated: May 8, 2024 - Viewed 25 days ago
The 'opportunity' table contains information about various business opportunities, it includes details such as the tap version, time of extraction, owner ID, and forecast category. This data can be used to track the progress of different opportunities, analyze their performance, and forecast future outcomes. It can also help to [Show more](#)
- sku_cost_lookup** .tbl - Table - main.prophet_forecast_schema - ryan.chynoweth@databricks.com - Updated: May 18, 2024

On the right side of the interface, there is a "Feedback on results" button and a navigation menu. The navigation menu includes: "All results", "Tables", "Notebooks", "Jobs", "Queries", "Dashboards", "Folders", "Git folders", "Endpoints", "Files", "Libraries", "Alerts", and "Experiments".

A red box highlights a dropdown menu in the top right corner, which is currently open. The dropdown menu shows "Sort by: Popularity" and a list of options: "Relevance" and "Popularity" (which is selected).

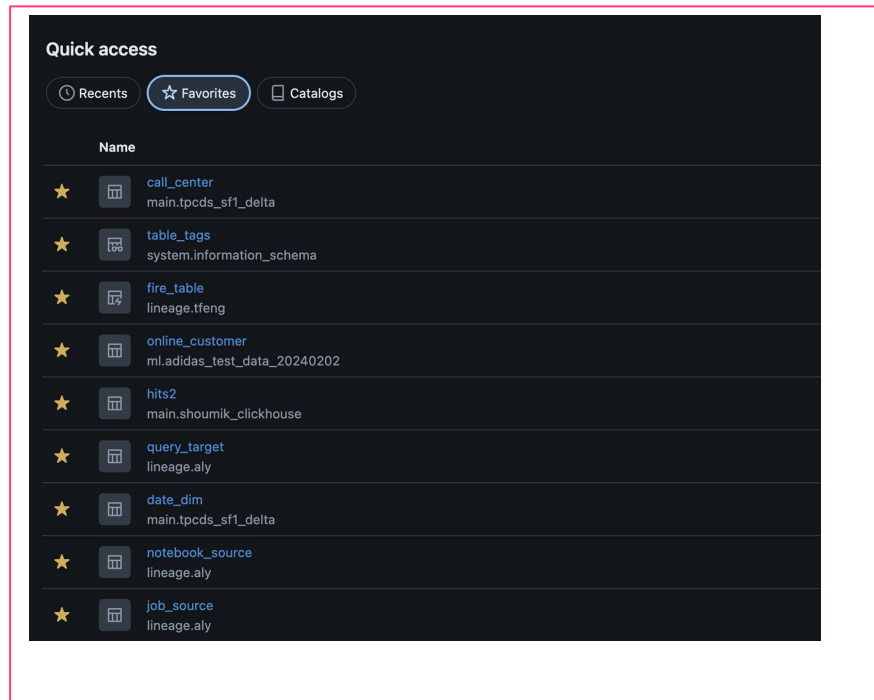
Schema Browser

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



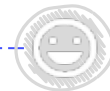
Catalog Explorer

- 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

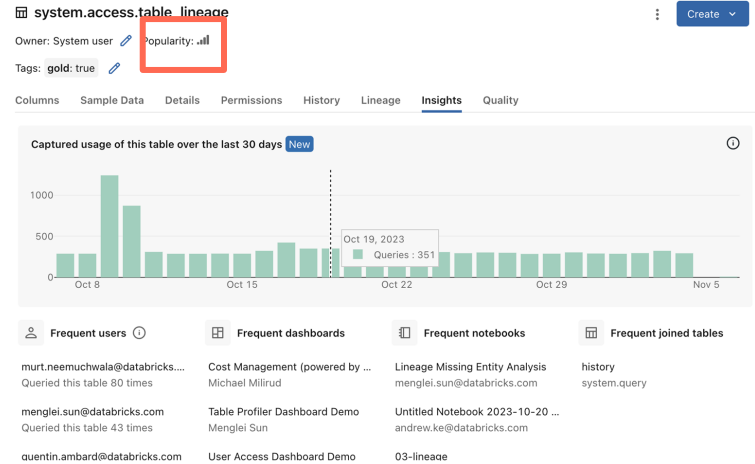


Explore data

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

About this table >

Owner: Tao Feng

Data source format: Delta

Last Updated: last year

Popularity:

Size: 677B, 1 file

Filter:

Tags:

AI Suggested Comment

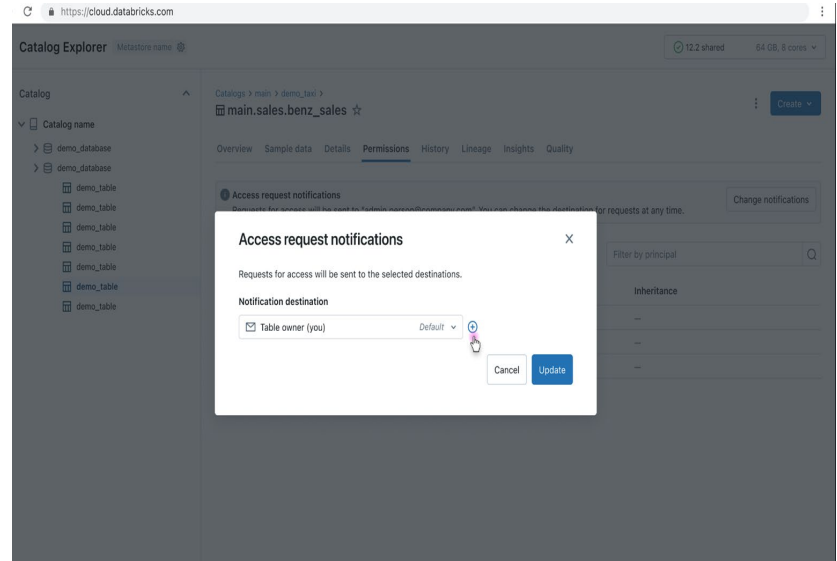
The 'notebook_source' table contains data about the sources of notebooks used in our analysis. It includes details about the two columns, c1 and c2, which are likely to be used as input variables for further analysis. This table can be useful for understanding the origin of notebooks and their potential impact on the results of our analysis. It can also help in tracking the sources of notebooks and their usage patterns, enabling better resource allocation and decision-making.

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

The screenshot displays the Databricks interface with four main sections: Frequent users, Frequent dashboards, Frequent notebooks, and Frequent joined tables. The 'Frequent queries' section is highlighted with a red box and contains the following SQL query:

```
select sum(usage_quantity) as dbu, usage_date from system.billing.usage b join distinct_ids t on b.record_id = t.record_id where u...
```

Below the highlighted query, there are two more queries, each starting with 'WITH q AS (select t.dbu - w.dbu as healthy_dbu, w.dbu as wasted_dbu, t.dbu_in_dollar - w.dbu_in_dollar as healthy_dbu_in_dollar, w...'.



End-to-End Lineage and Insight

Lineage Today

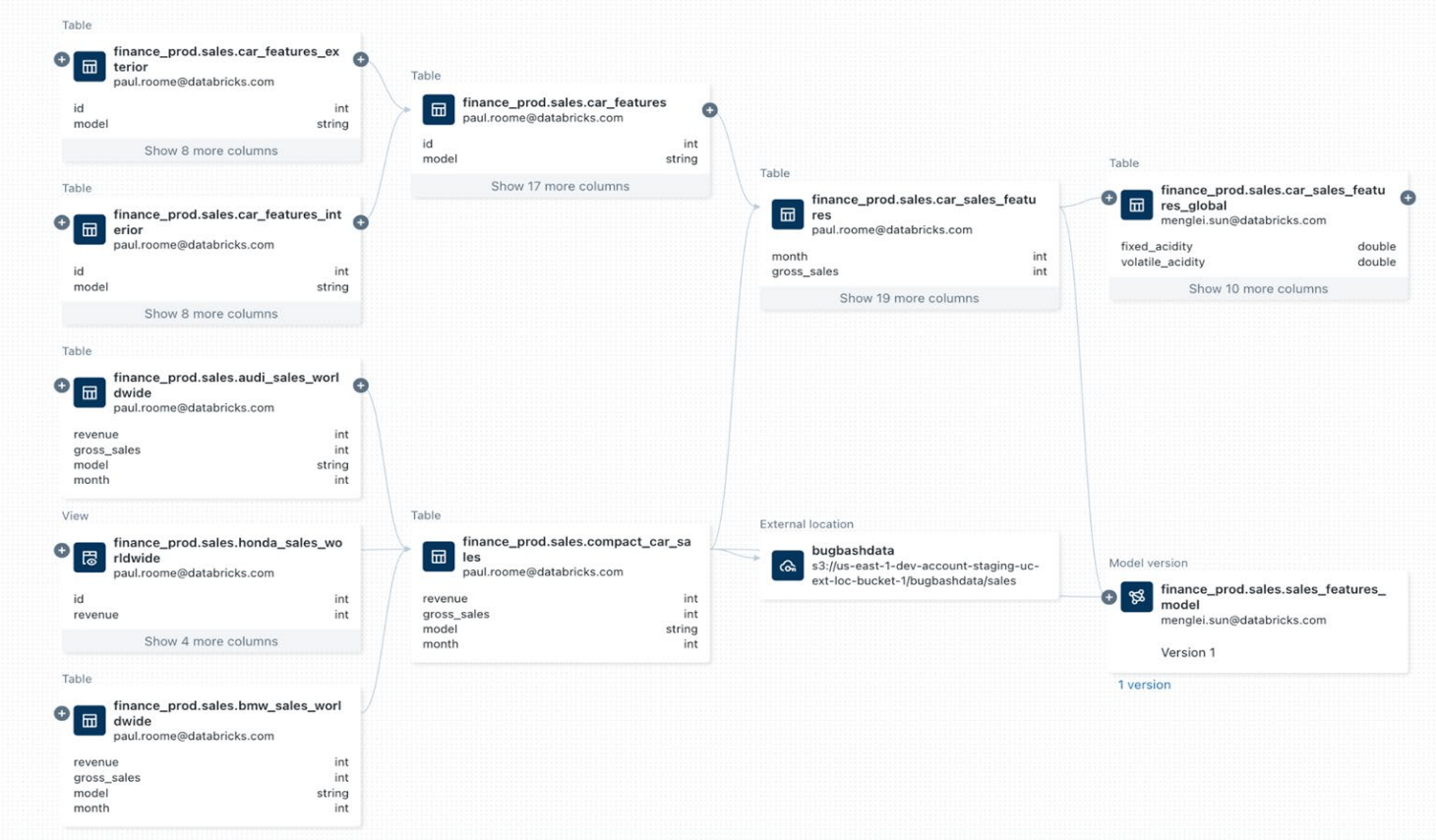
Automatic Capturing

- 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 !

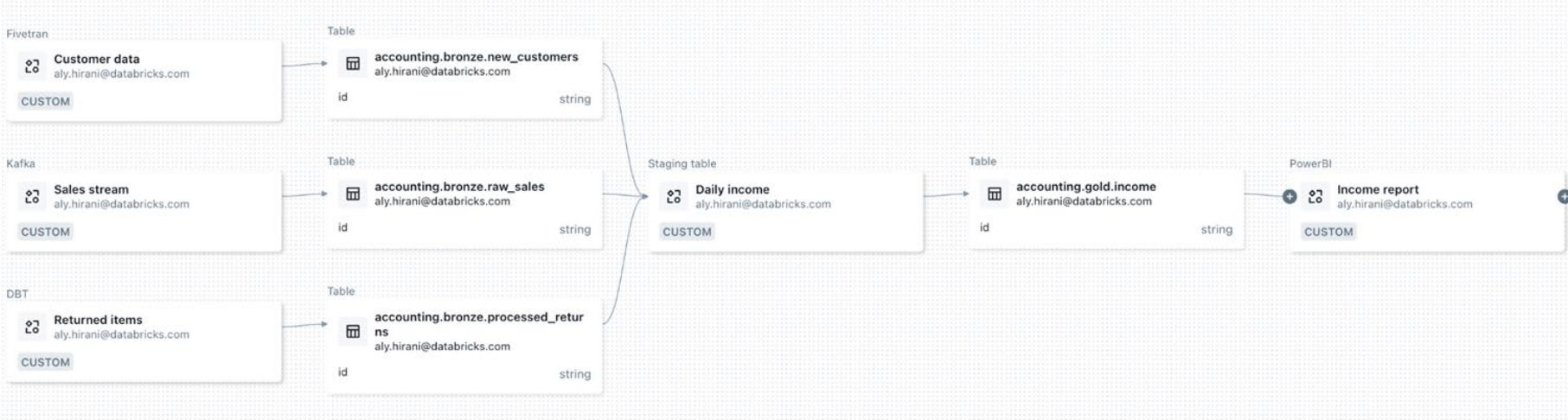
Before...



Example

Entity	Relationship
<pre>{ "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\"}" }</pre>	<pre>{ "source": { "provider_type": "CUSTOM", "guid": "kafka-customers" }, "target": { "provider_type": "DATABRICKS", "databricks_type": "TABLE", "guid": "sales.bronze.customers_raw" } }</pre>

After!



Next Steps

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