The Evolution of Data Catalogs

and Why They Matter

Alex Noonan

Developer Advocate

@alexnoonan6

@C00ldudeNoonan

Join us on Slack: dagster.io/slack in the #dagster-deep-dives channel

Agenda

What is a Data Catalog?

A Quick History of Data Catalogs

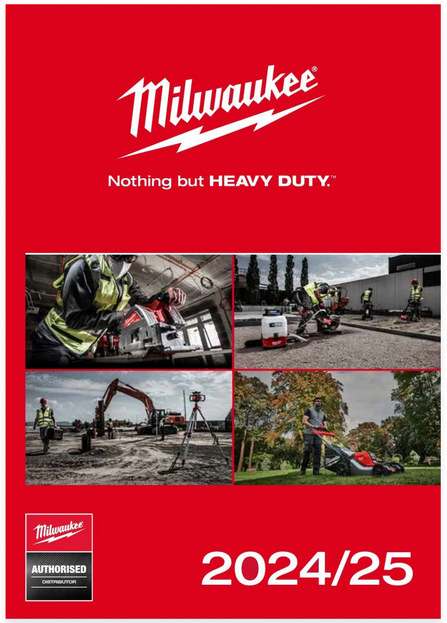
The Unfulfilled Promise

Dagster’s Approach

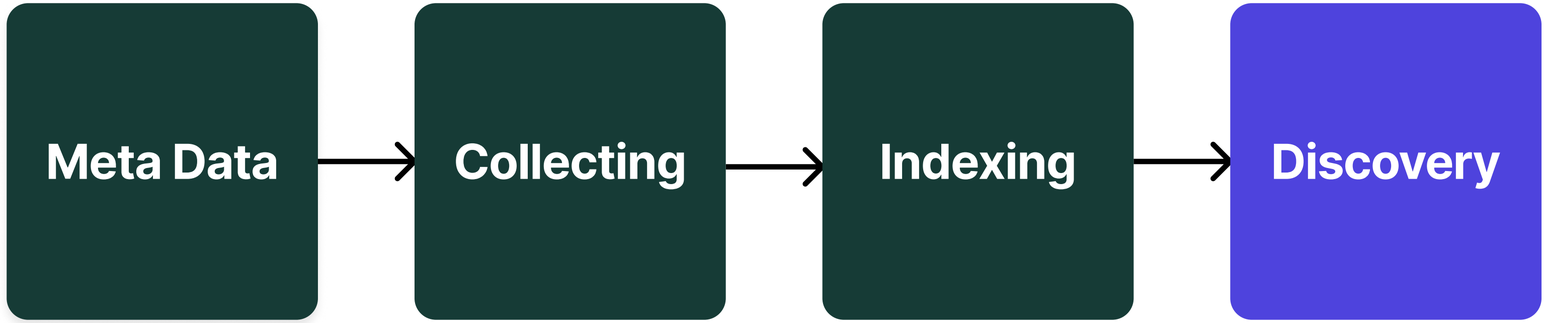
Catalog Demo

Q&A

What is a Data Catalog?



Its like any other catalog
but focused on Data



Metadata is the core input
to fuel data discovery

Timeline of Data Catalogs

3000 B.C	700 A.D	1876	1976	2018
Cuneiform Tablets	Scriptoriums	Dewey Decimal System	Relational Databases	Dagster

Clay Tablets in Ancient Mesopotamia

Durable

Standardization

Proto-Metadata & Indexing

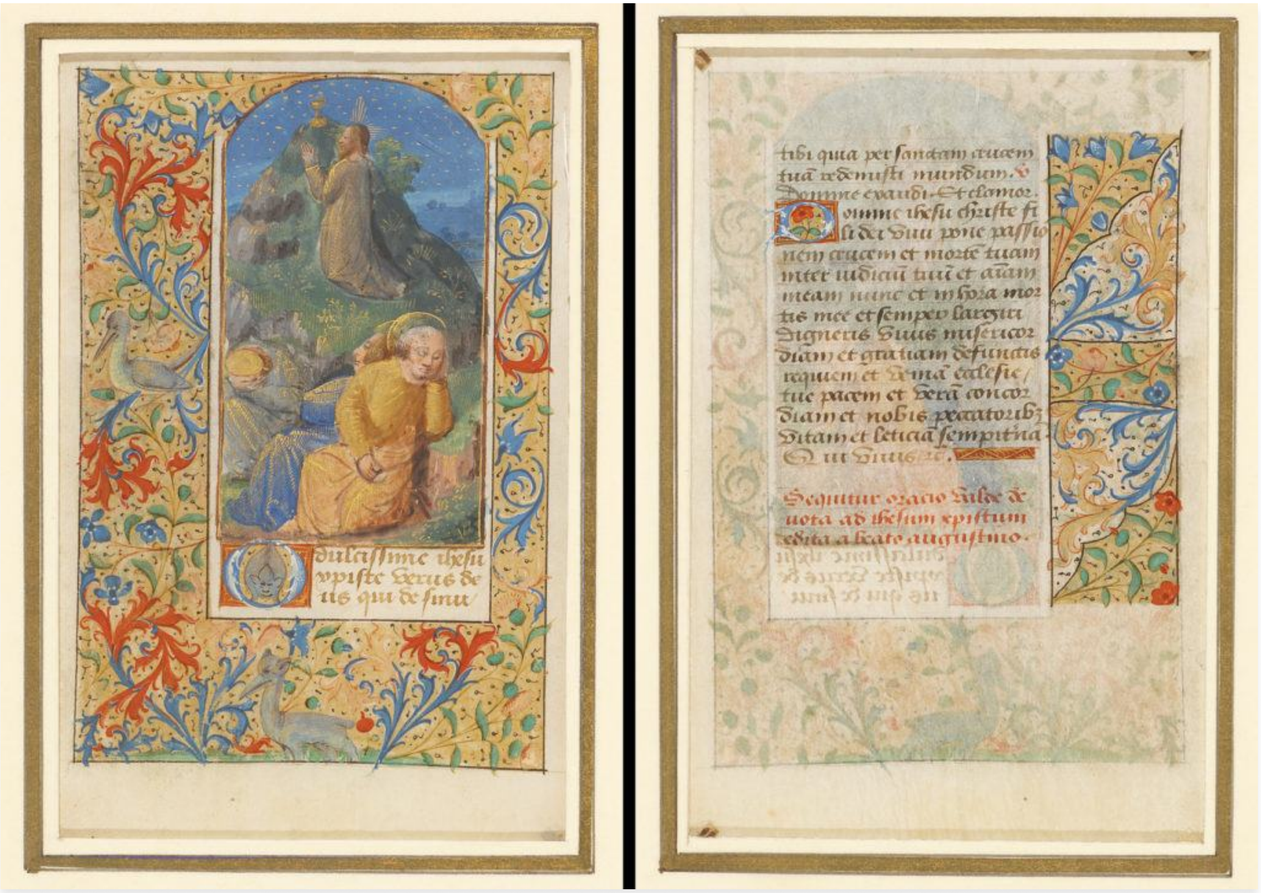
Receipt of wool from Drehem
Iraq circa 2200 B.C

Medieval Monks

Crudely grouping texts together

Catalogs physically maintained organized

Illumination



Dewey Decimal System

Standardization

Numerical System

Hierarchical

Dewey Numbers	Dewey Main Category	Books in this Category Answer this Question
000 - 099	General Information	How do we organize information?
100 - 199	Philosophy & Psychology	Who am I?
200 - 299	Religion	How did we get here?
300 - 399	Social Sciences	Who are the people around me?
400 - 499	Language	How can I communicate with others?
500 - 599	Science	How can I explain the world around me?
600 - 699	Technology	How can I control the world around me?
700 - 799	Arts & Recreation	How can I enjoy my free time?
800 - 899	Literature	What are the stories of our lives?
900 - 999	History & Geography	What was the world like in the past? What is it like now?

Relational Databases

Information Demands Standardization

Numerical System

Hierarchical

The Unfulfilled Promise

“I think I had a dashboard for this”

Where is Everything?

Its never been easier to create data assets, but that cuts both ways.

Lacking Intent

When Catalogs have only the metadata exhaust of data assets they lack the intent of the platform engineer.

Documentation Debt

Many Catalogs on the market require additional configuration and maintenance which is often neglected.

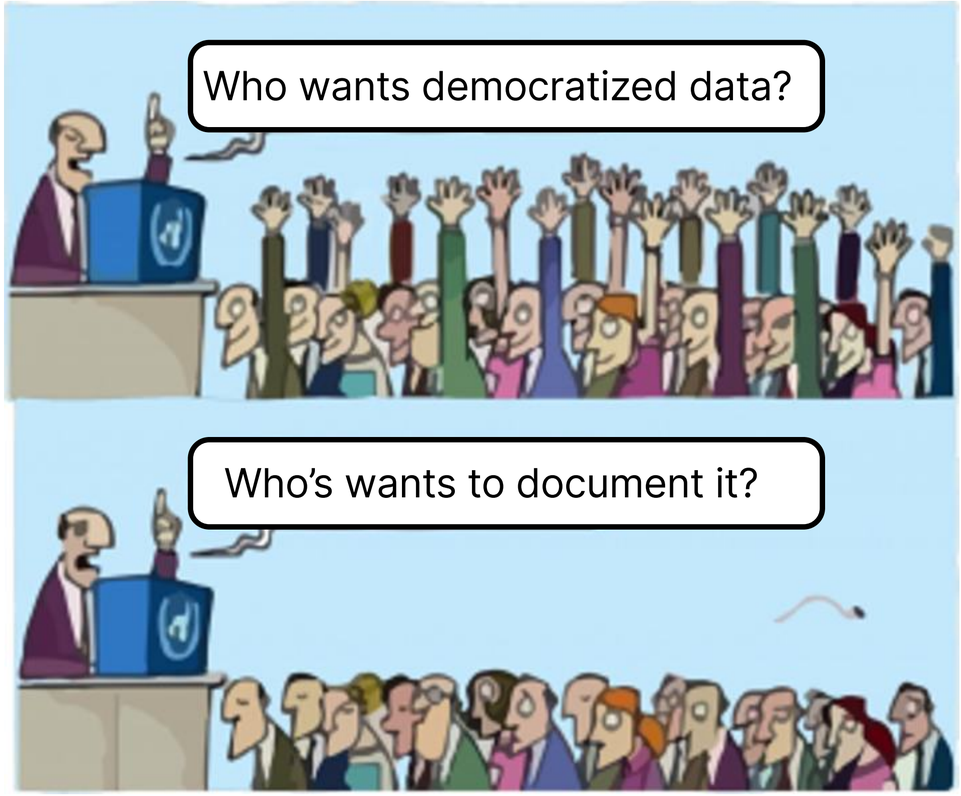
Data Governance We Want

Collaborative Stewardship

Data Lineage

Comprehensive Audit trials

Removing Data Silos

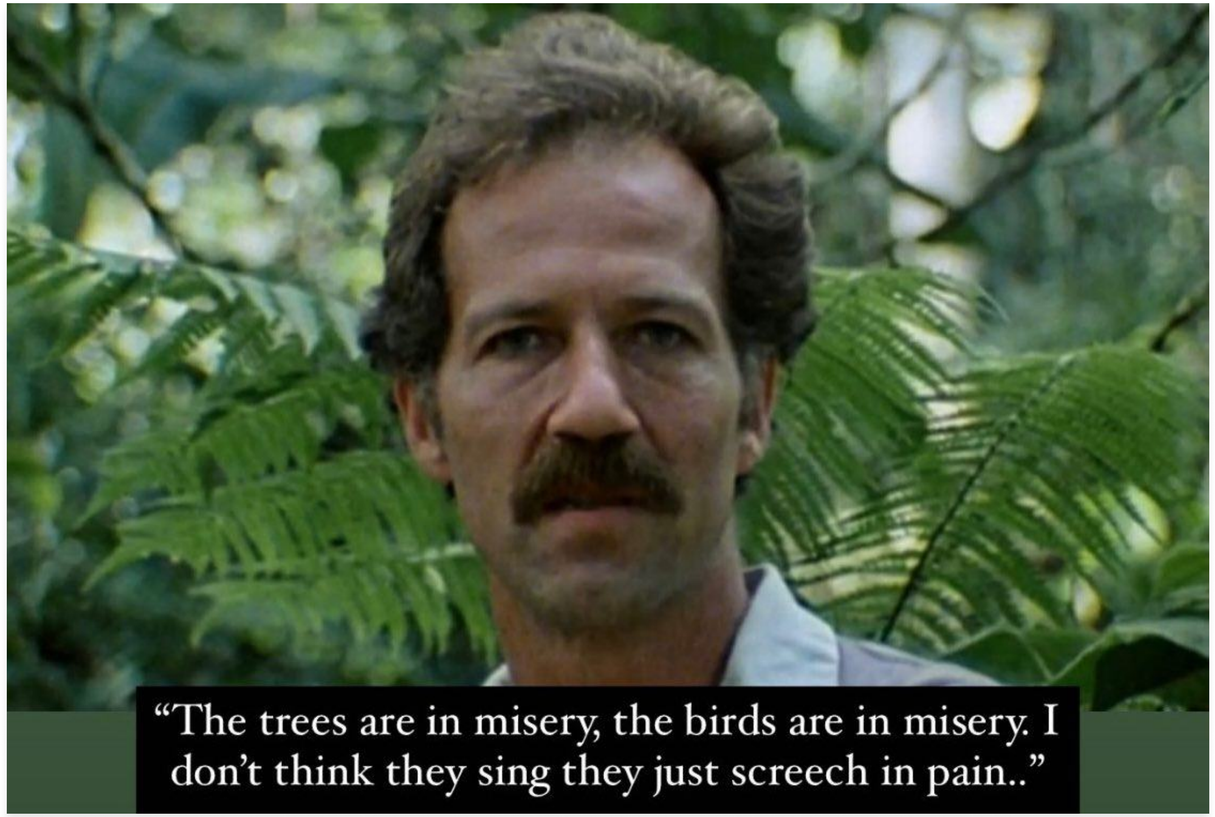


Documentation Debt

Tribal Knowledge

Hidden Costs

Data Discovery



“Senior Engineer describing how the data platform actually works”

Documentation as Code

Source of Truth

Flexibility

Automation

```
@asset(
    partitions_def=product_skus,
    io_manager_key="model_io_manager",
    key_prefix="MARKETING",
    ins={"sku_stats": AssetIn(key_prefix=["ANALYTICS"])},
    tags={
        **build_kind_tag("hex"),
        **build_kind_tag("s3"),
        },
)
def key_product_deepdive(context, sku_stats):
    """Creates a file for a BI tool based on
    the current quarters top product
    , represented as a dynamic partition"""
    key_sku = context.partition_key
    sku = sku_stats[sku_stats["sku"] == key_sku]
    context.add_output_metadata(
        {"sku_preview": MetadataValue.md(sku.head().to_markdown())}
    )
    return sku
```

Our Approach

Fighting Tool Fatigue

Switching Costs

Management Overhead

One more tool might fix me
(it wont)

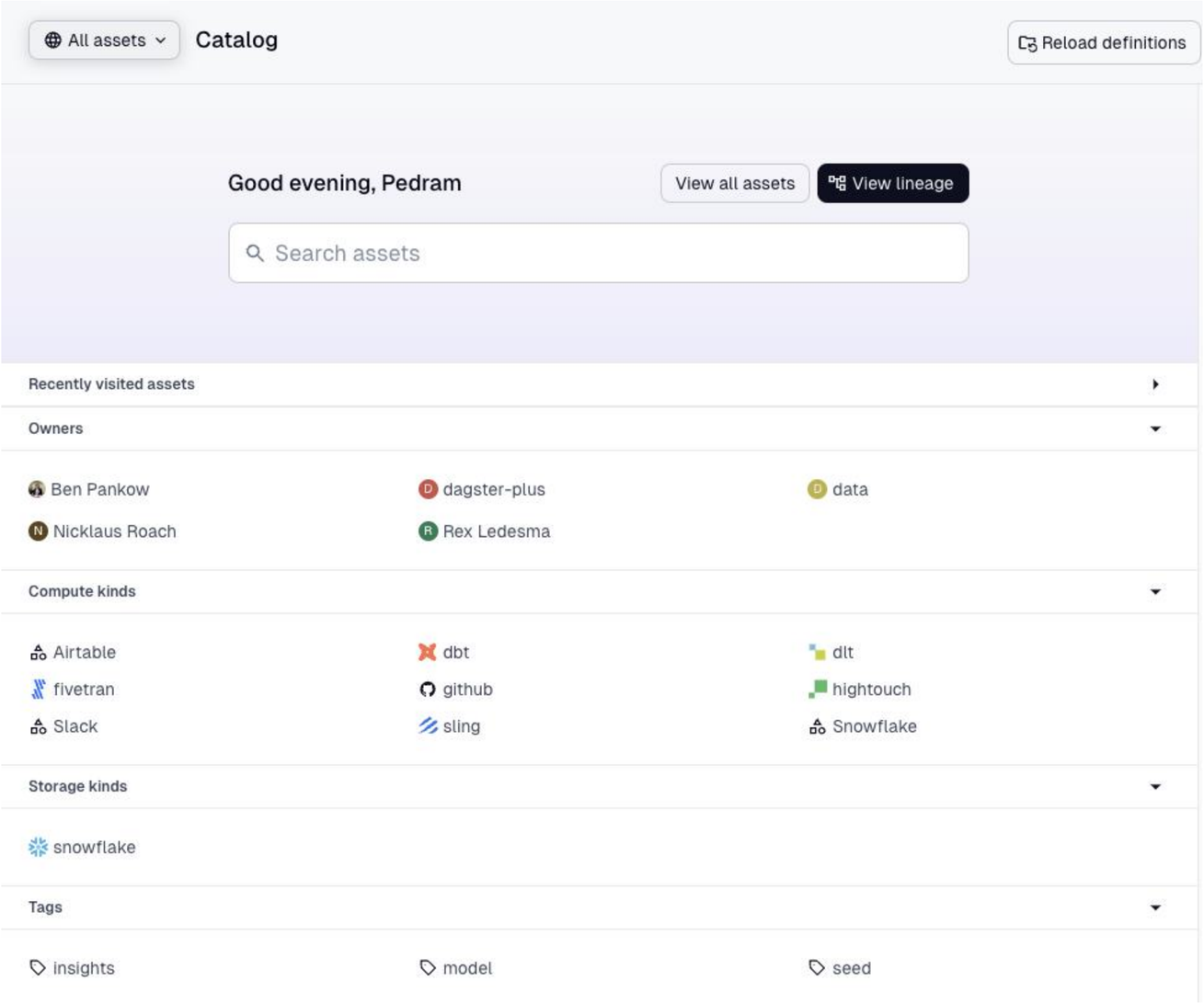
Developers, Developers, Developers

Software Defined Assets

Minimizing Overhead

Onboarding New Devs

Data Catalog Demo



Search by asset name, compute kind, group, owner, tag and more!

Wrapping Up

Data Platform System of Record

A built-in catalog for data platform teams that want a holistic view of their production data assets.

Developer Experience

No additional engineering hours required. The catalog is generated from your software defined assets and reflects your intention as a data platform engineer.

Better Data Discovery

Intuitive user experience for platform builders get what they need when they need it.

Next Steps & Resources

Join the Dagster Slack

Connect with other data practitioners. Share knowledge or find help. Use #ask-ai when you get stuck

dagster.io/slack

Sign up for Dagster+

Sign up for Dagster+ and get started with a free 30 day trial.

dagster.io

Follow us on Social

Never miss an update!

@dagster

@Dagster Labs

@dagsterio

@dagster-io