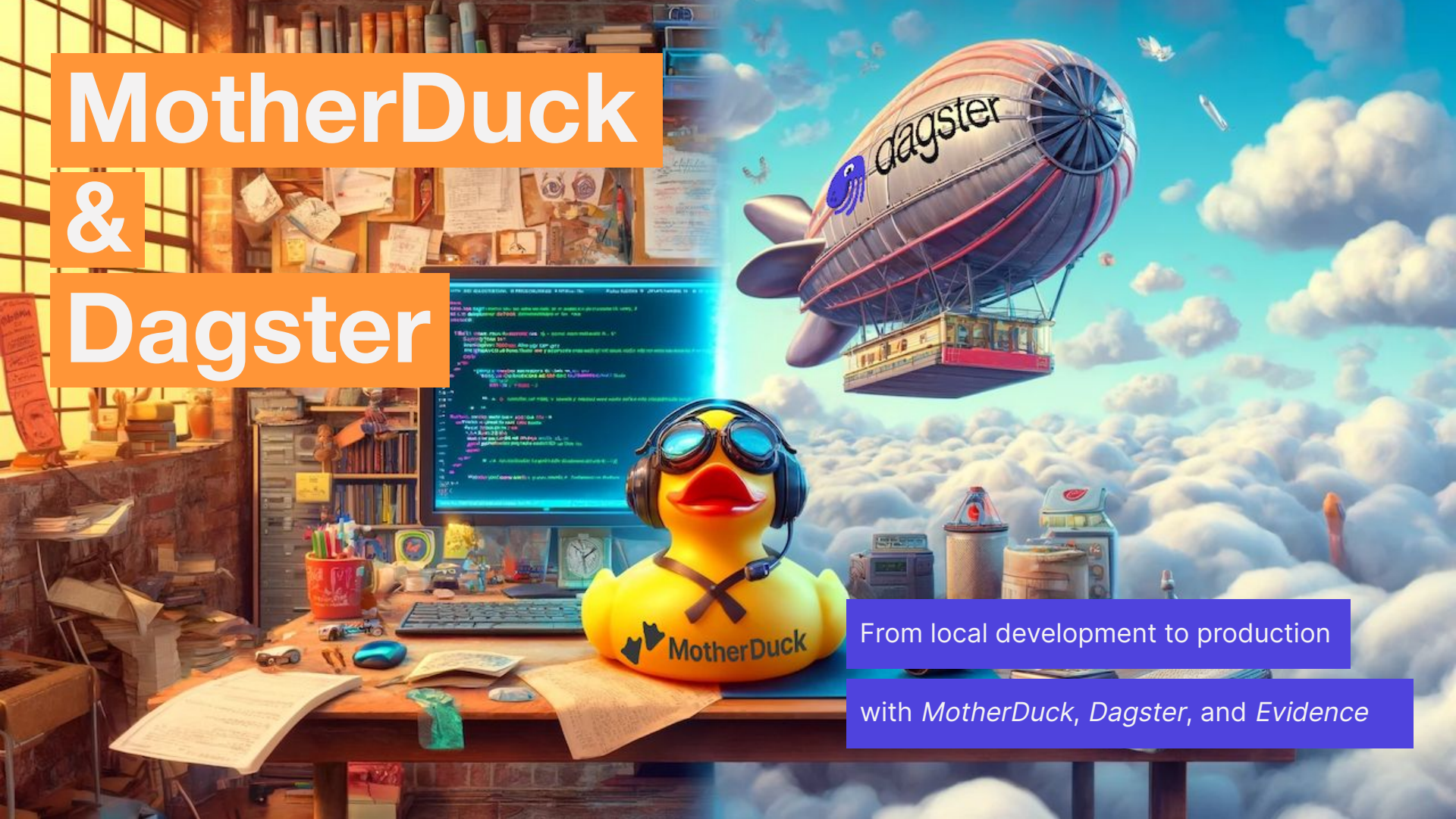
MotherDuck & Dagster



From local development to production

with *MotherDuck*, *Dagster*, and *Evidence*

Colton Padden

Data Engineer & Developer Advocate
Dagster Labs

@coltonpadden

Alex Monahan

Forward Deployed Software Engineer
MotherDuck & DuckDB Labs

@__AlexMonahan__

Problem Statement

When building data pipelines, it can be difficult to efficiently work in a local environment and transition to production.

Local Data Engineering Complexities

- Difficult to mock data sources
- Difficult to write unit tests
- Scale of data can be difficult to run locally
  - Dual query execution for the win!
- Integration with systems

The Solution

Using tools that emphasize developer productivity like Dagster, MotherDuck, and Evidence make it simple to develop locally, and ship to production with confidence.

The Stack

1. Dagster for orchestration
2. DuckDB / MotherDuck as a data warehouse
3. dbt for transformations
4. Evidence for reporting and visualizations

https://dagster.io

Dagster is a Python data orchestrator that allows you to build a world-class data platform, with an emphasis on developer workflow and productivity.

Has officially supported integrations for a variety of toolchains including DuckDB, MotherDuck, and dbt.

https://evidence.dev

Evidence is a lightweight platform for building data applications and dashboards. It is powered by DuckDB WASM and has integrations for many data warehouses including MotherDuck.

Reports are intuitively written in markdown with embedded SQL blocks, and there is a large library of visualizations out of the box!

https://motherduck.com

MotherDuck is a serverless data warehouse powered by DuckDB, a lightweight, in-process SQL Analytics Engine that is taking the data world by storm. It supports hybrid, dual engine query processing to easily leverage the compute power of the devices at our fingertips.

Because DuckDB can run anywhere, we can run the same code locally as we run in MotherDuck for fast local development and seamless deployment.

Let’s get into it.

QUESTIONS

1. Why rethink the cloud data warehouse?

2. What is MotherDuck?

3. What experiences become ducking awesome?

Why rethink the cloud data warehouse?

BIG DATA IS DEAD

LAPTOPS ARE POWERFUL

SERVERS ARE ENORMOUS

EVERYONE LIKES DUCKDB

BIG DATA IS DEAD
LONG LIVE EASY DATA

Big Data is just marketing

- **BigQuery**: ~95% < 1TB
- **SingleStore**: ~80% smallest instance
- **Gartner**: Most EDWs < 100GB
- **A16Z**: B2B < 1TB
- **A16Z**: B2C < 10TB

If data size is not most important…

BIG DATA IS DEAD
…MAKE DATA EASY

LAPTOPS ARE POWERFUL

SERVERS ARE ENORMOUS

EVERYONE LIKES DUCKDB

MotherDuck

DUCKDB MAKES ANALYTICS EASY

A great burger is more than just good beef!

Easy to install / no dependencies
Run anywhere (including the browser)
Query dataframes directly
Friendliest SQL syntax in the world

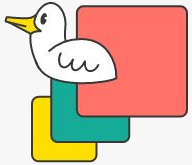
MotherDuck - Making Data Warehouses Personal

I - Git Operations for Your Database

II - Serverless: Only pay for CPU/disk you use

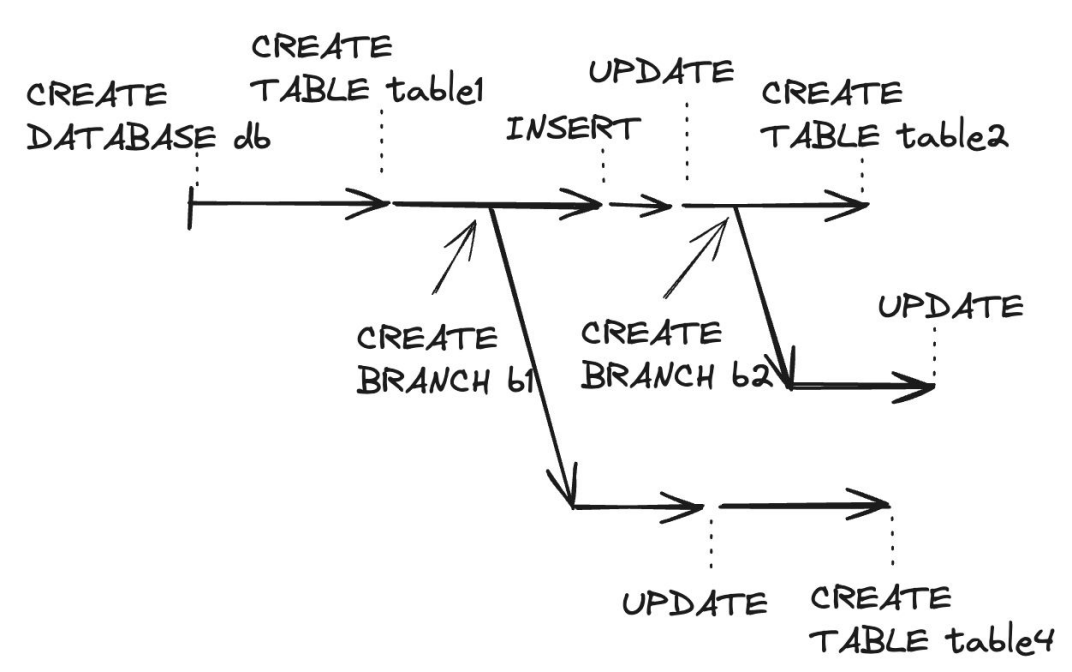
III - Pull from Postgres & MySQL with just dbt

IV - Local development, powered by DuckDB



Git Operations for Your Database

- Today: Zero-copy clone for worry free development
- Tomorrow: Branch, merge, diff, rollback, time travel



If local compute is powerful…

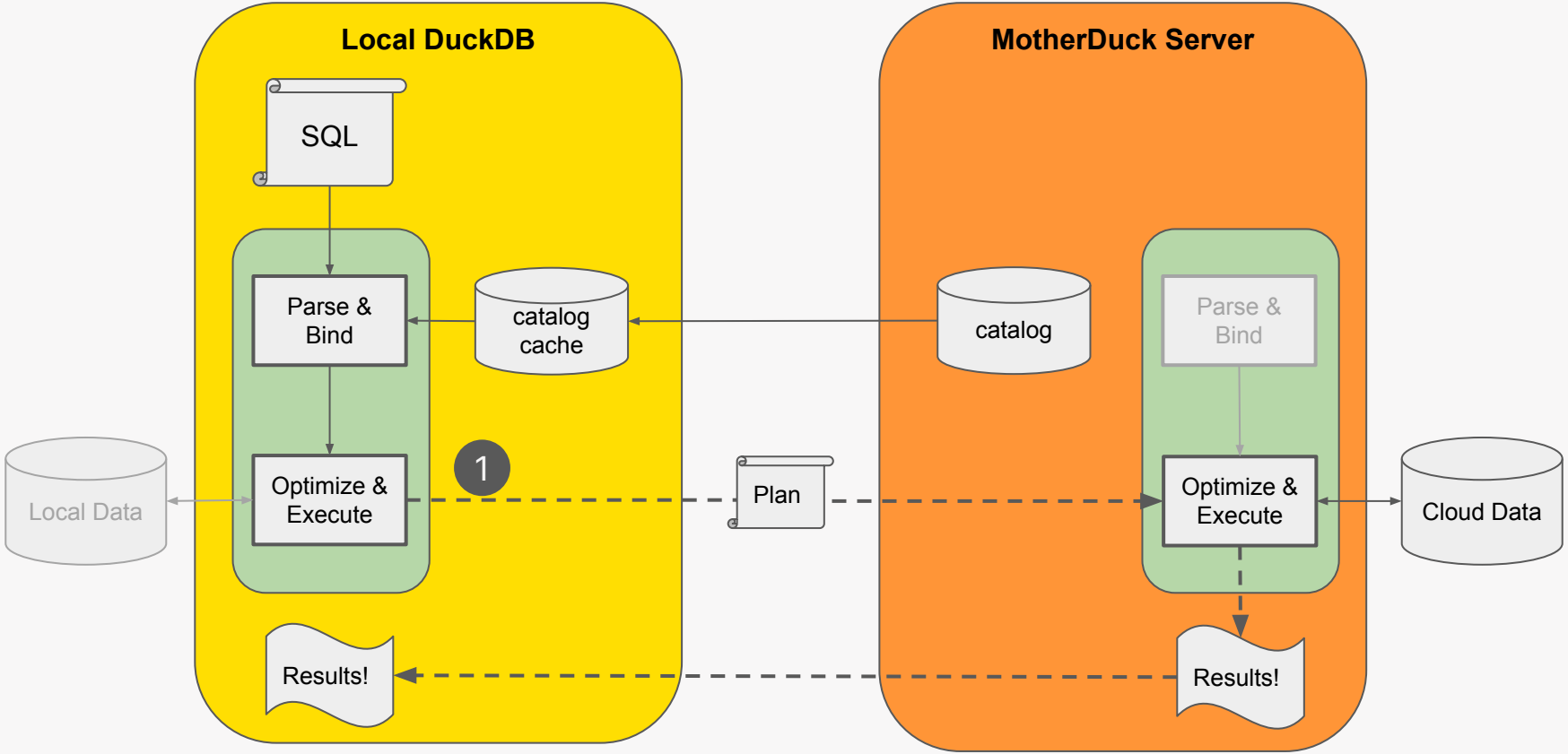
BIG DATA IS DEAD
…MAKE DATA EASY

LAPTOPS ARE POWERFUL
…MOVE WORK LOCALLY

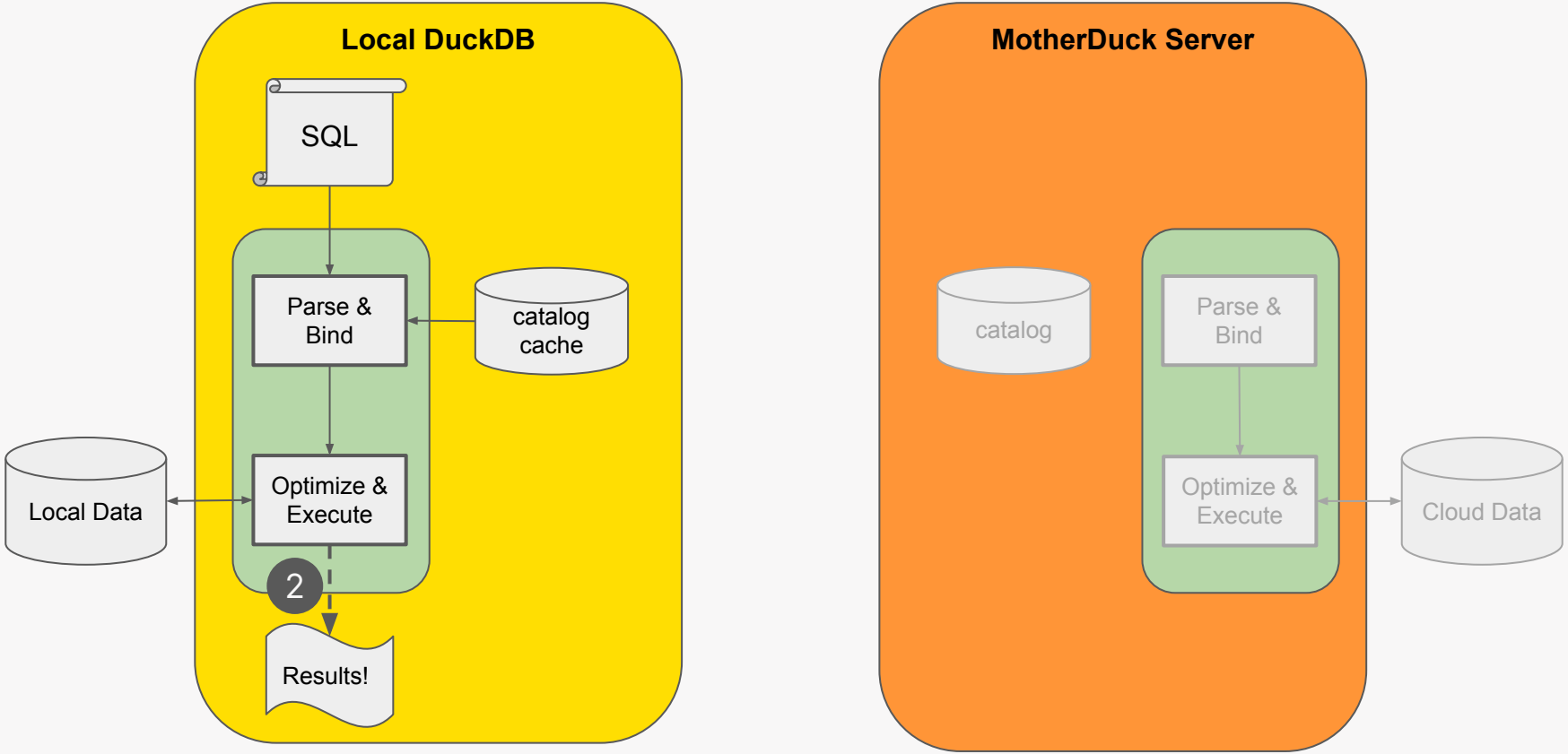
SERVERS ARE ENORMOUS

EVERYONE LIKES DUCKDB

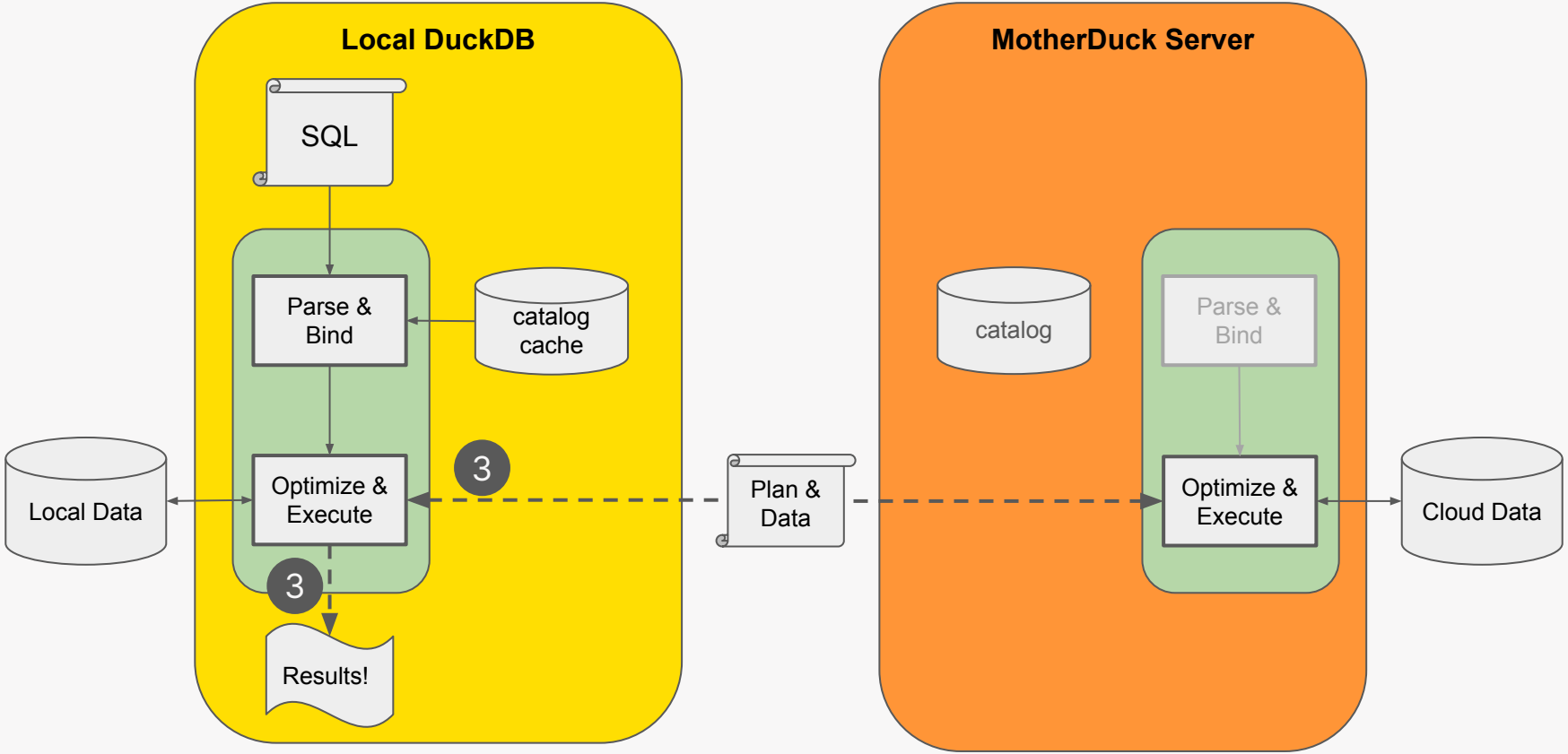
Dual Engine Execution: Cloud data only



Dual Engine Execution: Local data only



Dual Engine Execution: Local and Cloud data



If cloud servers are enormous…

BIG DATA IS DEAD
…MAKE DATA EASY

LAPTOPS ARE POWERFUL
…MOVE WORK LOCALLY

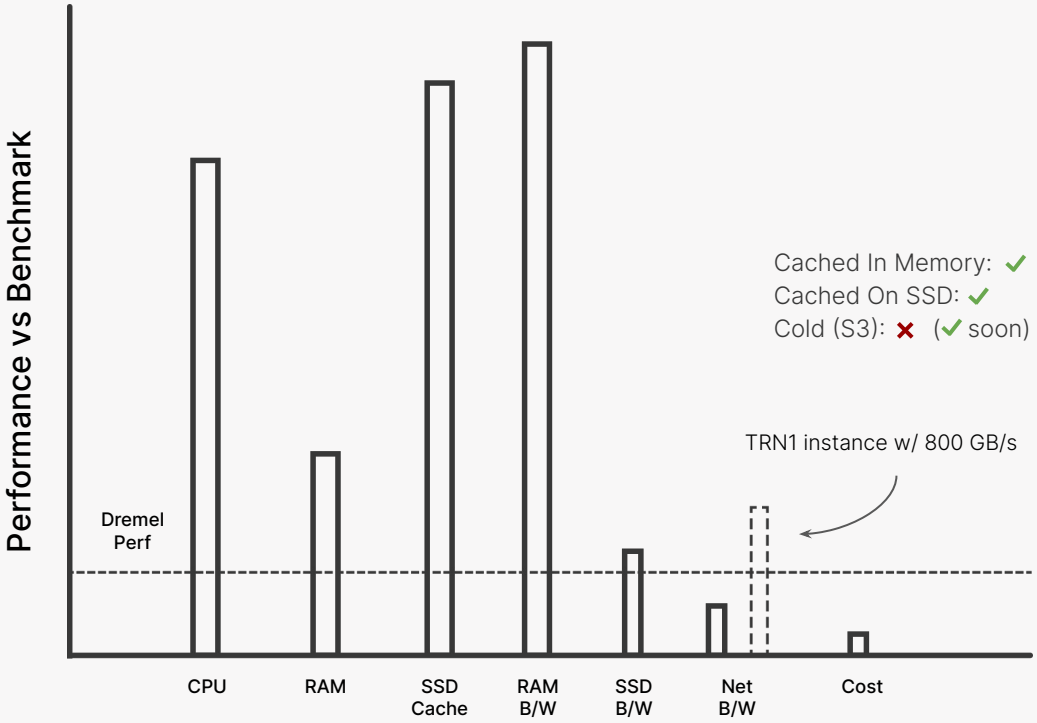
SERVERS ARE ENORMOUS
…LET’S SCALE UP

EVERYONE LIKES DUCKDB

DREMEL IN A BOX

2010: 3,000 nodes

2024: 1 AWS I4i.metal node
- 128 CPU
- 1 TB RAM
- $11/hour

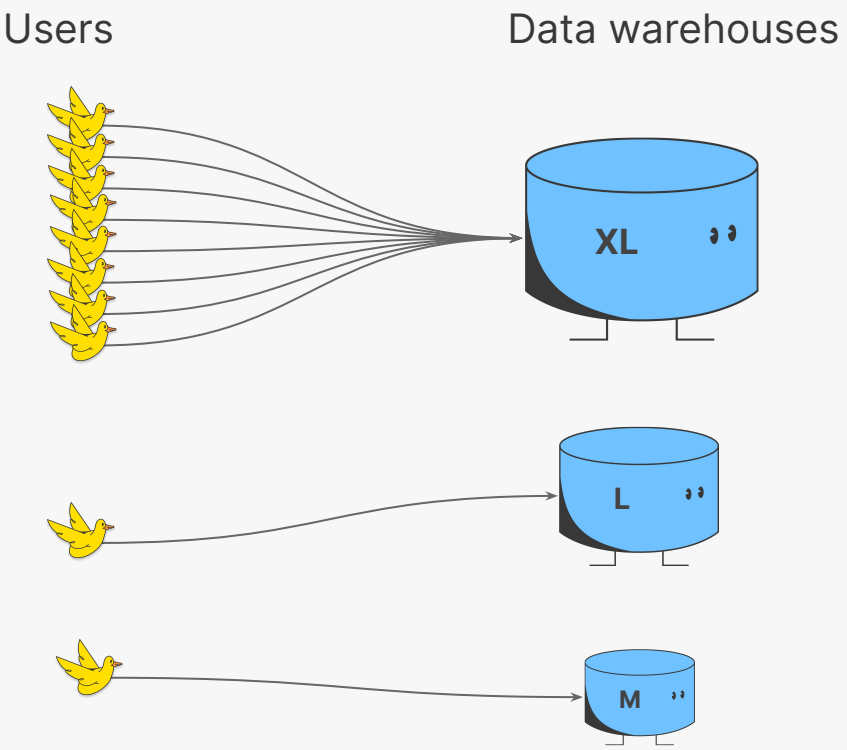


Dremel: Interactive Analysis of Web-Scale Datasets

Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, Theo Vassilakis
Google, Inc.

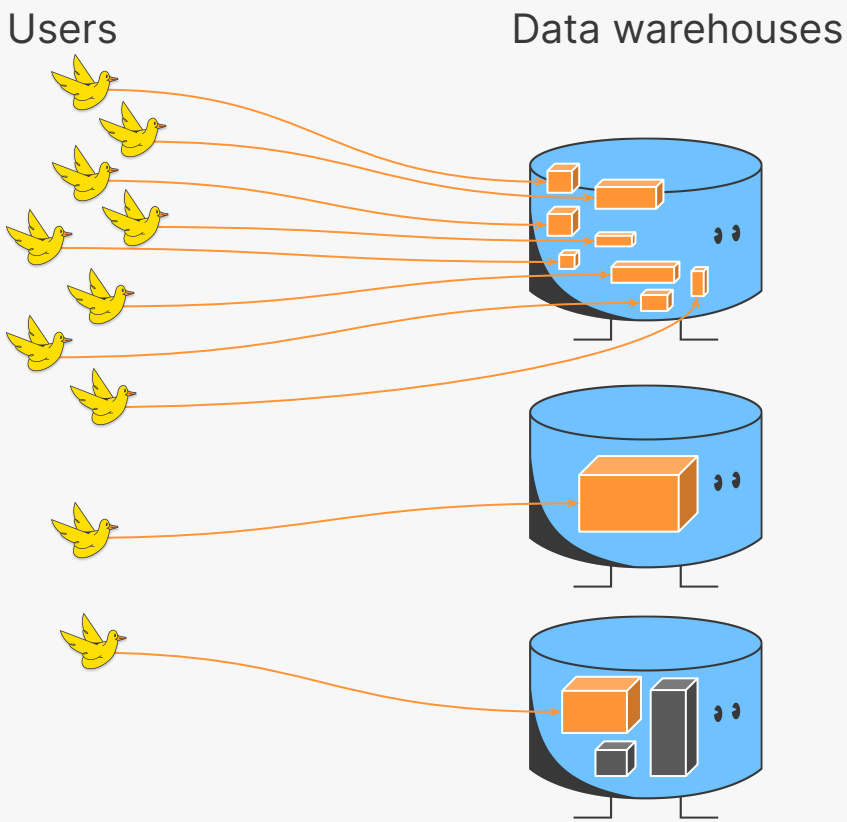
Industry Standard: Bin-Pack The Workloads

- Bin-pack or isolate?
- Re-balance? Resize?
- Cost attribution?
- Usage auditing?
- Cost/perf optimization?



MotherDuck Vision: The Personal Data Warehouse

- Each user gets their own efficient, isolated serverless runtime (“duckling”)
- Choice to bin-pack or isolate is MotherDuck’s problem, not users
- Easy cost attribution, usage auditing, and optimizations



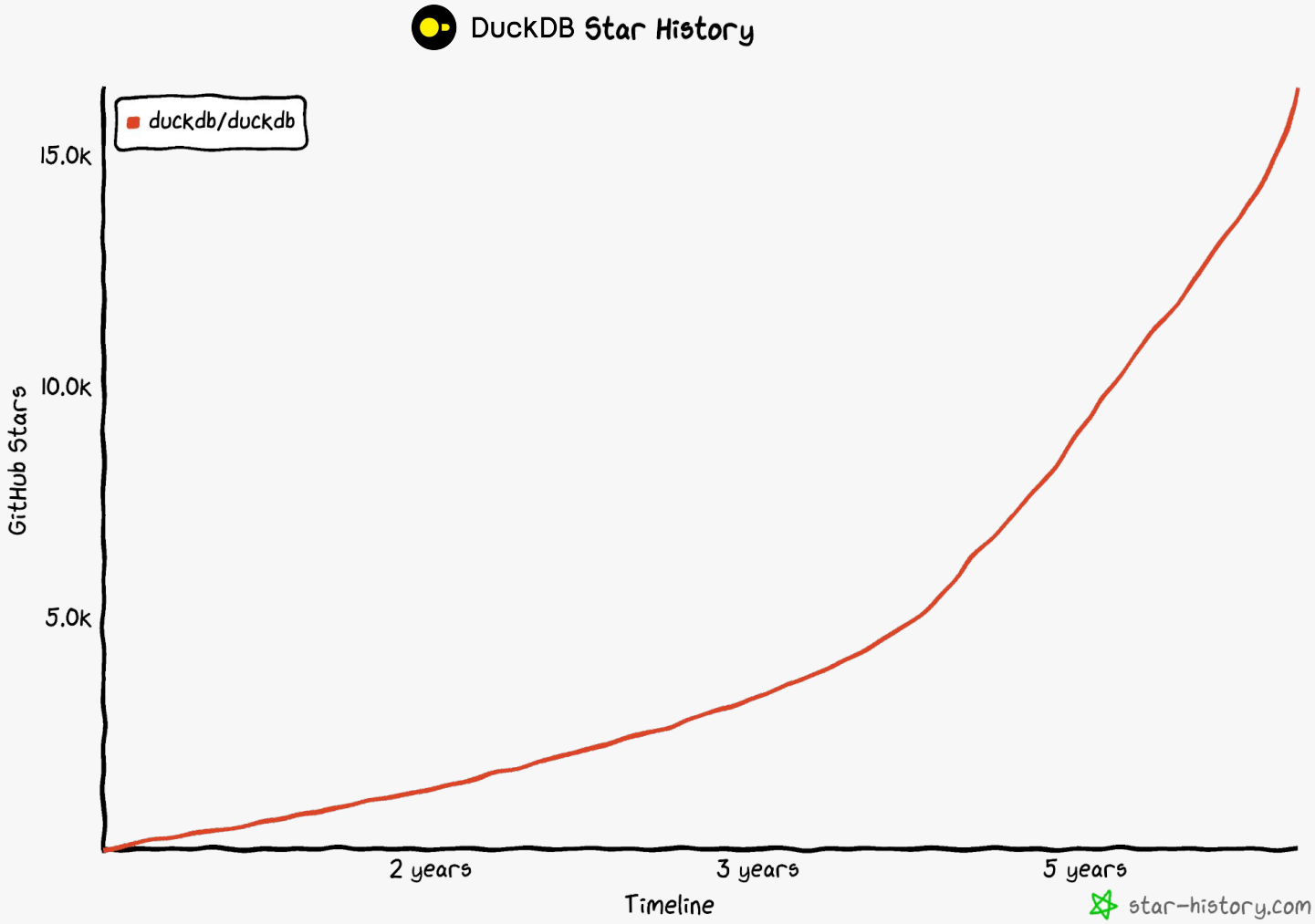
If everyone loves DuckDB…

BIG DATA IS DEAD
…MAKE DATA EASY

LAPTOPS ARE POWERFUL
…MOVE WORK LOCALLY

SERVERS ARE ENORMOUS
…LET’S SCALE UP

EVERYONE LIKES DUCKDB
…LET’S SUPERCHARGE IT!


DuckDB Star History
duckdb/duckdb
15.0k
10.0k
5.0k
GitHub Stars
2 years
3 years
5 years
Timeline
star-history.com

DuckDB at a glance

Simple

DuckDB is easy to install and deploy. It has zero external dependencies and runs in-process in its host application or as a single binary.

Read more →

Portable

DuckDB runs on Linux, macOS, Windows, and all popular hardware architectures. It has idiomatic client APIs for major programming languages.

Read more →

Feature-rich

DuckDB offers a rich SQL dialect. It can read and write file formats such as CSV, Parquet, and JSON, to and from the local file system and remote endpoints such as S3 buckets.

Read more →

Fast

DuckDB runs analytical queries at blazing speed thanks to its columnar engine, which supports parallel execution and can process larger-than-memory workloads.

Read more →

Extensible

DuckDB is extensible by third-party features such as new data types, functions, file formats and new SQL syntax.

Read more →

Free

DuckDB and its core extensions are open-source under the permissive MIT License.

Read more →

Extending DuckDB with MotherDuck

I - Access from multiple processes

II - Share with your team

III - Scale with the cloud

IV - Run DuckDB everywhere: ELT, dbt, & BI

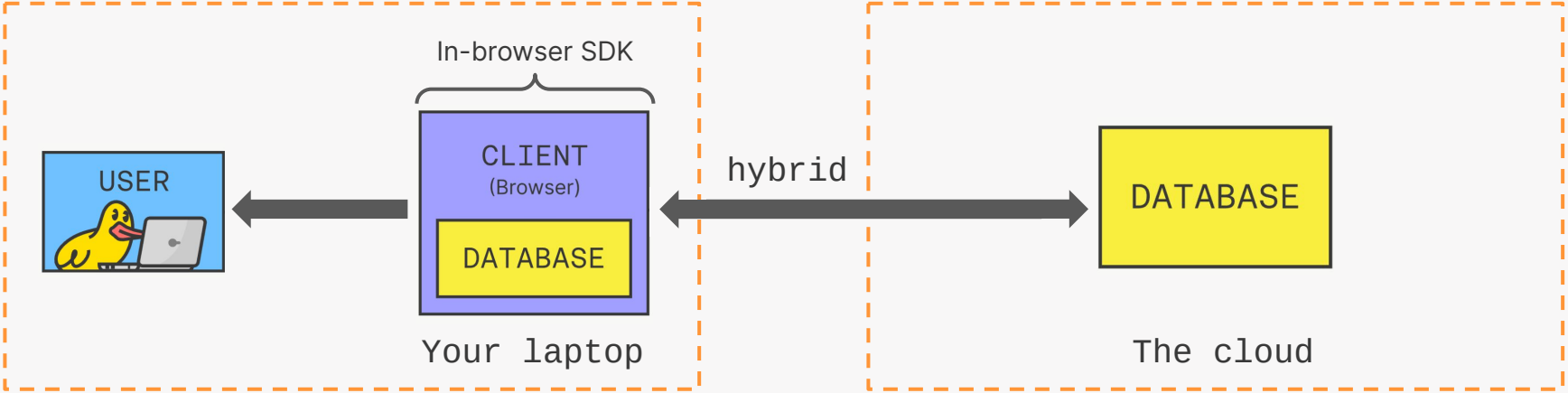
…Now what becomes ducking awesome?

DATA EXPLORATION AT THE SPEED OF THOUGHT

DATA TEAMS DEVELOP LOCALLY THEN PUSH TO THE CLOUD

DATA APPS GET BLAZING SPEED AND RICH INTERACTIVITY

MotherDuck - In-Browser Data Applications



- MotherDuck's WASM SDK brings analytics to the browser
- 60-frames-per-second, interactive in-browser experiences
- Faster, cheaper, and better than state-of-the-art

Data App Demo!

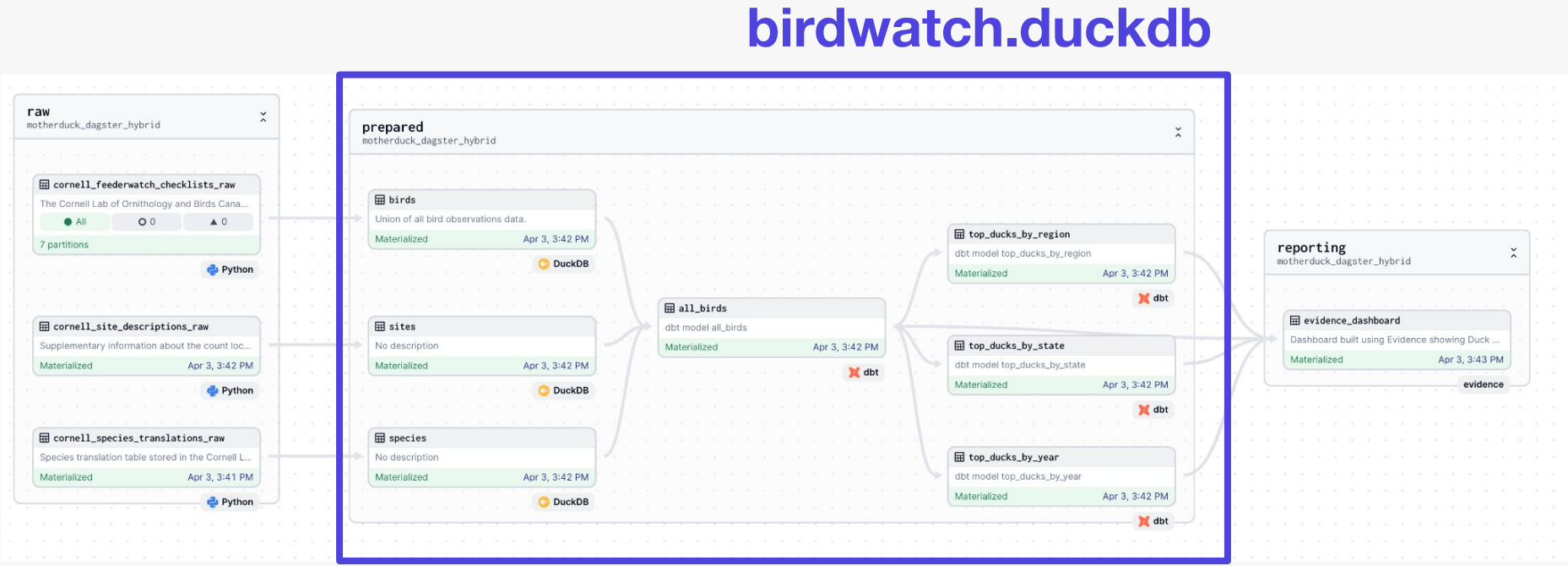
https://motherduckdb.github.io/wasm-client/mosaic-integration/

Example Project Walkthrough

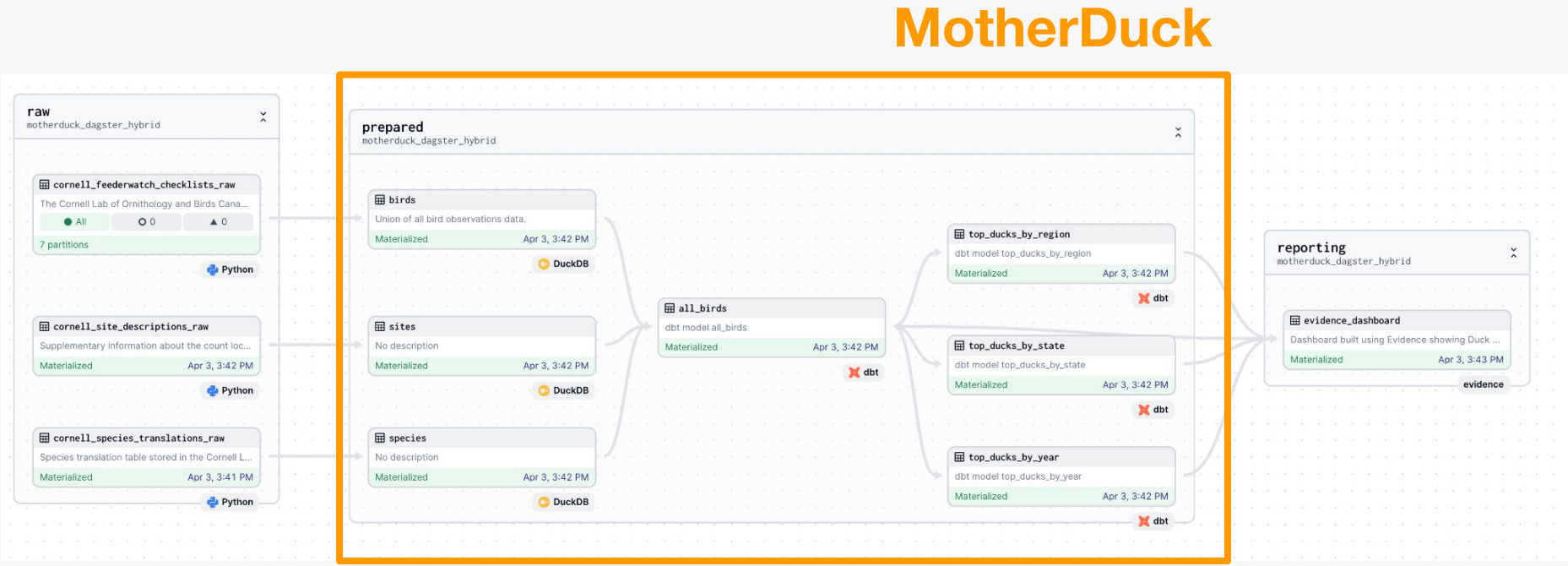
The Code

bit.ly/dagster-motherduck-demo

The Pipeline [Development]



The Pipeline [Production]



The End Product (via Evidence)

evidence

Home

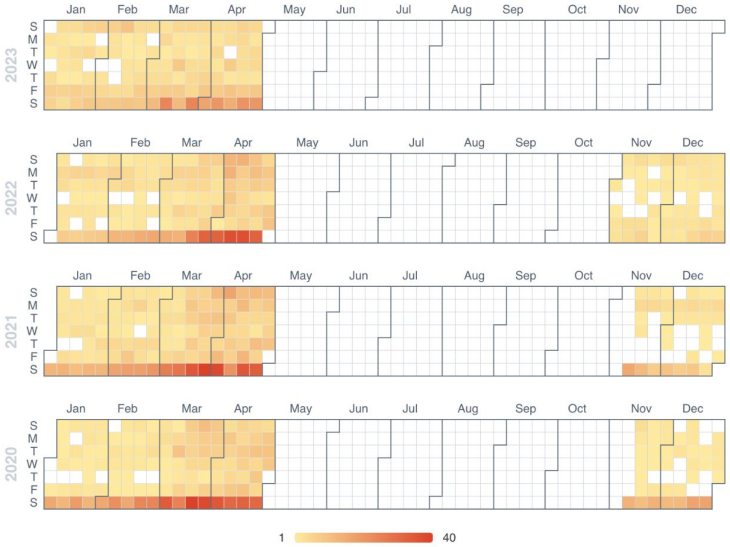
Duck Observations

There were **15** checklists, **20** observations, and **49** total species of Ducks.

Most observations occur on Saturdays – the FeederWatch dataset is seasonal, operating from November to April of each year.

Calendar Heatmap

Daily Sales

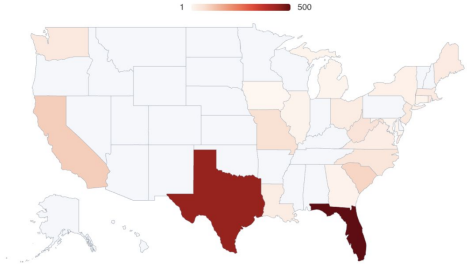


On this page

The Duck Details
Project Feederwatch
Dagster Asset Lineage
Metrics

U.S. State Heatmap

By far, the most duck observations occur in Florida, either these people have a lot of ducks or enough time on their hands to report them to Feeder Watch.



Project Structure

```
$ tree -L 2
.
├── Makefile
├── README.md
├── data
│   └── raw
├── dbt_project
│   ├── dbt_packages
│   ├── dbt_project.yml
│   ├── logs
│   ├── models
│   ├── profiles.yml
│   ├── snapshots
│   ├── target
│   └── tests
├── motherduck_dagster_hybrid
│   ├── __init__.py
│   ├── __pycache__
│   ├── assets.py
│   ├── constants.py
│   └── resources.py
├── pyproject.toml
├── reports
│   ├── README.md
│   ├── evidence.plugins.yaml
│   ├── node_modules
│   ├── package-lock.json
│   ├── package.json
│   ├── pages
│   ├── sources
│   └── static
├── setup.cfg
└── setup.py
```

Brief Code Overview

Resources

resources.py

```python
from dagster_duckdb import DuckDBResource

duckdb_resource = DuckDBResource(
    database=EnvVar("DUCKDB_DATABASE"),
)
```

.env

```
# MotherDuck / DuckDb Connection String
#
# Can either be a simple <filename>.duckdb, or a connection string to MotherDuck.
#
# References:
# - https://motherduck.com/docs/authenticating-to-motherduck/
DUCKDB_DATABASE="md:dagster_hybrid_demo?motherduck_token=xxx
```

This is where the magic happens

Assets

assets.py

- Assets are defined as normal Python functions with an asset decorator
- Reusable resources are accessible to the asset function as a parameter
- Metadata can be returned from the asset and viewed in the Dagster UI

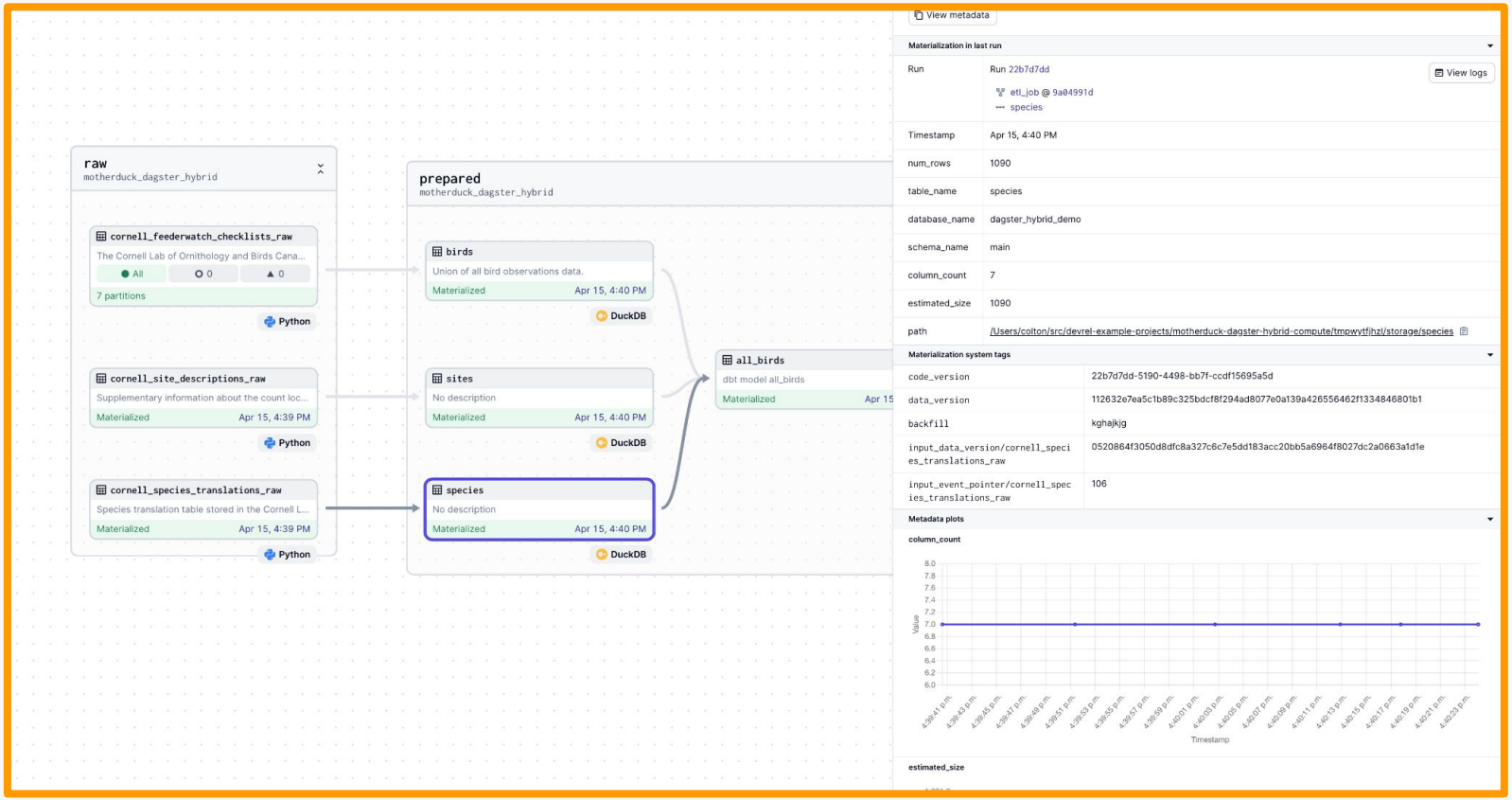
```python
@asset(
    deps=[cornell_species_translations_raw],
    compute_kind="duckdb",
    group_name="prepared",
)
def species(context: AssetExecutionContext, duckdb: DuckDBResource):
    species_csv_path = file_relative_path(__file__, SPECIES_TRANSLATION_FPATH)
    context.log.info("Loading species file: %s", species_csv_path)
    with duckdb.get_connection() as conn:
        conn.execute(f"""
            create or replace table species as (
                select * from read_csv_auto('{species_csv_path}')
            )
        """)

        nrows = conn.execute("""
            select count(*) from species
        """).fetchone()

        metadata = conn.execute("""
            select * from duckdb_tables() where table_name = 'species'
        """).pl()

    return MaterializeResult(
        metadata={
            "num_rows": nrows[0] if nrows else 0,
            "table_name": metadata["table_name"][0],
            "database_name": metadata["database_name"][0],
            "schema_name": metadata["schema_name"][0],
            "column_count": metadata["column_count"][0],
            "estimated_size": metadata["estimated_size"][0],
        }
    )
```

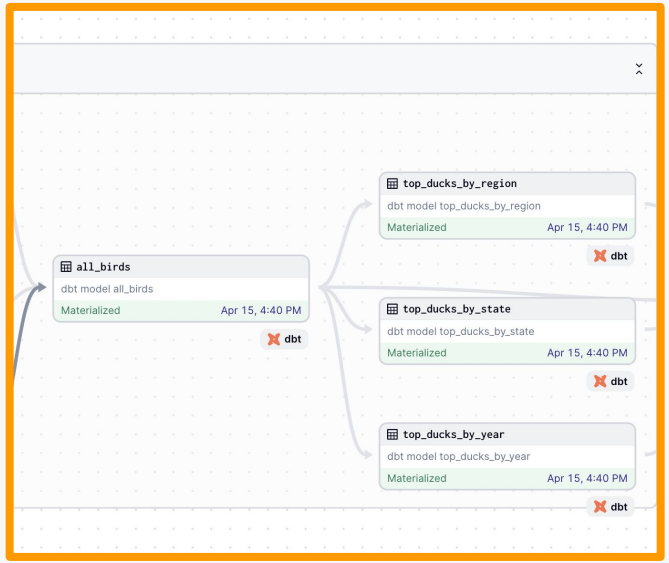
Assets


raw
motherduck_dagster_hybrid
cornell_feederwatch_checklists_raw
The Cornell Lab of Ornithology and Birds Cana...
All
0
0
7 partitions
Python
cornell_site_descriptions_raw
Supplementary information about the count loc...
Materialized
Apr 15, 4:39 PM
Python
cornell_species_translations_raw
Species translation table stored in the Cornell L...
Materialized
Apr 15, 4:39 PM
Python
prepared
motherduck_dagster_hybrid
birds
Union of all bird observations data.
Materialized
Apr 15, 4:40 PM
DuckDB
sites
No description
Materialized
Apr 15, 4:40 PM
DuckDB
species
No description
Materialized
Apr 15, 4:40 PM
DuckDB
all_birds
dbt model all_birds
Materialized
View metadata
Materialization in last run
Run
Run 22b7d7dd
View logs
etl_job @ 9a04991d
species
Timestamp
Apr 15, 4:40 PM
num_rows
1090
table_name
species
database_name
dagster_hybrid_demo
schema_name
main
column_count
7
estimated_size
1090
path
/Users/colton/src/devrel-example-projects/motherduck-dagster-hybrid-compute/tmpwytfjhzl/storage/species
Materialization system tags
code_version
22b7d7dd-5190-4498-bb7f-ccdf15695a5d
data_version
112632e7ea5c1b89c325bdcf8f294ad8077e0a139a426556462f1334846801b1
backfill
kghajkjg
input_data_version/cornell_species_translations_raw
0520864f3050d8dfc8a327c6c7e5dd183acc20bb5a6964f8027dc2a0663a1d1e
input_event_pointer/cornell_species_translations_raw
106
Metadata plots
column_count
Value
Timestamp
estimated_size

dbt

```python
dbt_manifest_path = dbt_parse_invocation.target_path.joinpath("manifest.json")

@dbt_assets(manifest=dbt_manifest_path, dagster_dbt_translator=CustomDagsterDbtTranslator())
def dbt_birds(context: OpExecutionContext, dbt: DbtCliResource):
    yield from dbt.cli(["build"], context=context).stream()
```




```
$ tree dbt_project/models/
dbt_project/models/
├── all_birds.sql
├── sources.yml
├── top_ducks_by_region.sql
├── top_ducks_by_state.sql
└── top_ducks_by_year.sql
```

dbt (continued)

models/top_ducks_by_region.sql

```sql
select

    country,
    region,
    bird_name,
    total_count

from (

    select

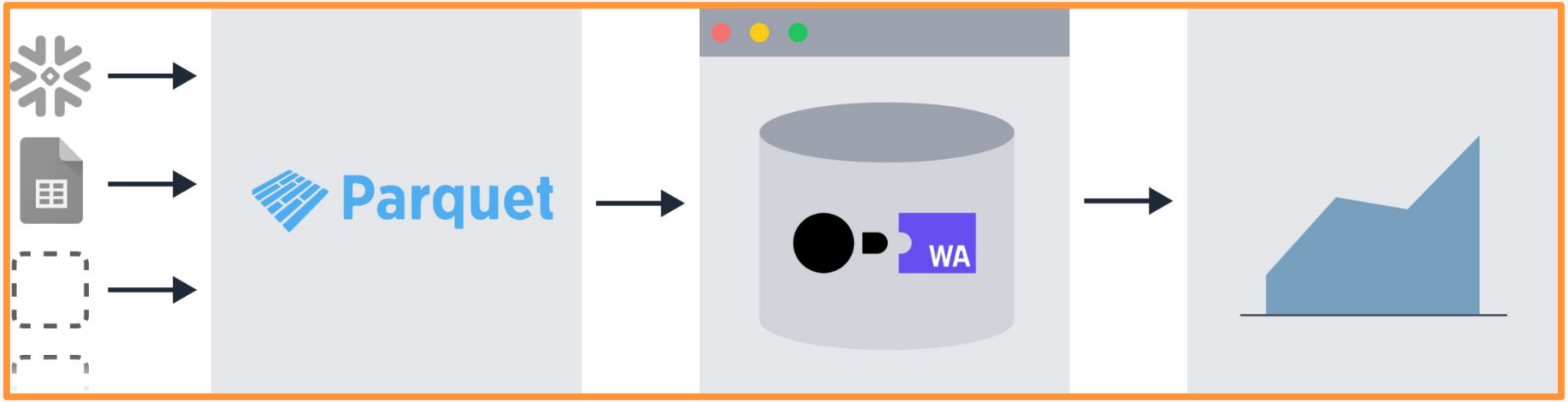
        country,
        region,
        bird_name,
        sum(species_count) as total_count,
        row_number() over (partition by region order by sum(species_count) desc) as rank

    from {{ ref('all_birds') }}
    where region in ('CA', 'ON', 'NY', 'PA', 'VA', 'FL')
    group by 1, 2, 3

) where rank <= 5
```

Evidence

- Evidence syncs source data to Parquet files
- We define our MotherDuck connection, and our sources
- Source tables are then referenced from our reports



Evidence (cont.)

pages/01-duck-observations.md

```
### Duck Observations


There were <strong><Value data={duck_counts} value='n_checklists' /></strong> checklists, <strong><Value data={duck_counts} value='n_observations' /></strong> observations, and <strong><Value data={duck_counts} value='total_species' /></strong> total species of Ducks.

Most observations occur on Saturdays -- the FeederWatch dataset is seasonal, operating from November to April of each year.

<CalendarHeatmap
    data={duck_counts_recent}
    date=obs_date
    value=n_observations
    title="Calendar Heatmap"
    subtitle="Daily Observations"
/>

```sql duck_counts_recent
select

    obs_date,
    date_trunc('year', obs_date) as obs_year,
    count(distinct checklist_id) as n_checklists,
    count(distinct observation_id) as n_observations,
    sum(species_count) as total_species

from dagster_hybrid_demo.all_ducks
where obs_date > '2020-01-01'
group by all
order by 1 desc
```
```

Evidence

assets.py

```python
@asset(compute_kind="evidence", group_name="reporting", deps=[dbt_birds])
def evidence_dashboard():
    """Dashboard built using Evidence showing Duck metrics."""
    evidence_project_path = file_relative_path(__file__, "../reports")
    subprocess.run(["npm", "--prefix", evidence_project_path, "install"])
    subprocess.run(["npm", "--prefix", evidence_project_path, "run", "sources"])
    subprocess.run(["npm", "--prefix", evidence_project_path, "run", "build"])
```

- Dashboard generation is declared as a Dagster asset
- Source queries are downloaded locally and cached as Parquet
- Static HTML / JavaScript web pages are built for high-performance dashboards leveraging DuckDB WASM

Least Observed Ducks

1. The **Long Tailed Duck**

> The long-tailed duck (Clangula hyemalis) or coween,[2] formerly known as the oldsquaw, is a medium-sized sea duck that breeds in the tundra and taiga regions of the arctic and winters along the northern coastlines of the Atlantic and Pacific Oceans. It is the only member of the genus Clangula. [more info]

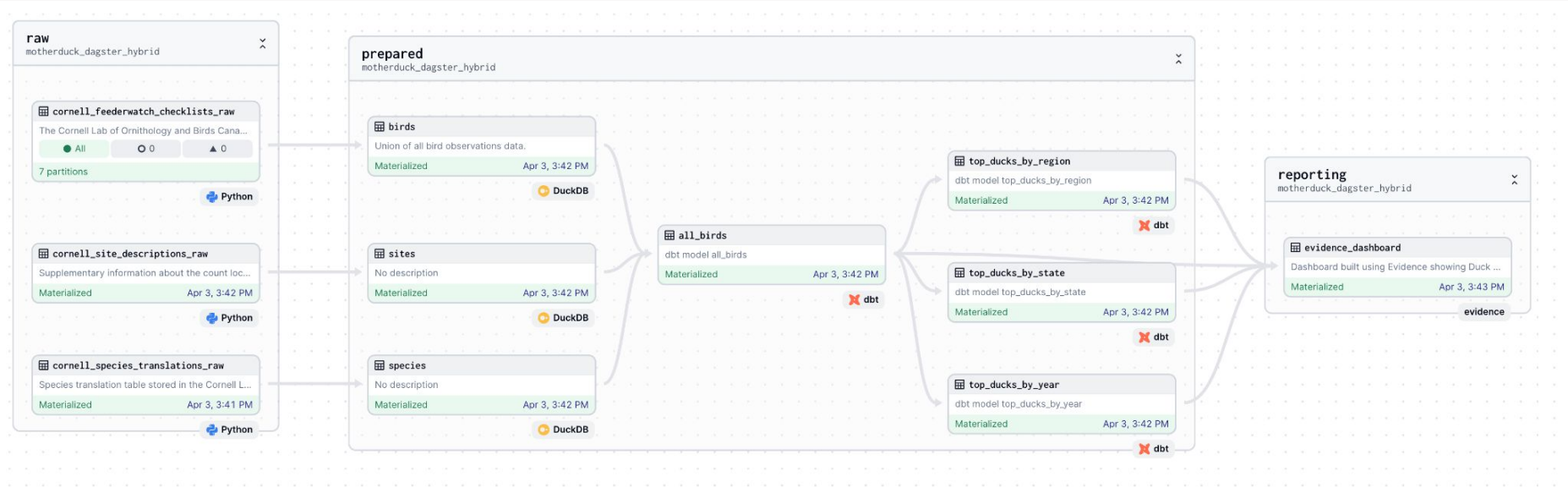
2. The **Mandarin Duck**

> The mandarin duck (Aix galericulata) is a perching duck species native to the East Palearctic. It is sexually dimorphic, males showing a dramatic difference from the females.[3] It is medium-sized, at 41–49 cm (16–19 in) long with a 65–75 cm (26–30 in) wingspan. It is closely related to the North American wood duck, the only other member of the genus Aix. 'Aix' is an Ancient Greek word which was used by Aristotle to refer to an unknown diving bird, and 'galericulata' is the Latin for a wig, derived from galerum, a cap or bonnet.[4] Outside of its native range, the mandarin duck has a large introduced population in the British Isles and Western Europe, with additional smaller introductions in North America. [more info]

3. The **Muscovy Duck**

> The Muscovy duck (Cairina moschata) is a duck native to the Americas, from the Rio Grande Valley of Texas and Mexico south to Argentina and Uruguay. Feral Muscovy ducks are found in New Zealand, Australia, and in Central and Eastern Europe. Small wild and feral breeding populations have also established themselves in the United States, particularly in Florida, Louisiana, Massachusetts, the Big Island of Hawaii, as well as in many other parts of North America, including southern Canada. [more info]

Pipeline Review



Additional Links & Resources

- Dagster DuckDB Integration
- MotherDuck Evidence Integration
- Evidence Core Concepts: Data Sources
- Project FeederWatch
- MotherDuck Docs

Q & A