PyData

Data Lineage: Where 'It Depends' Finally Gets an Answer

2024-11-28

Tomas Peluritis

Head of Data @ Mediatech

AKA Uncle Data

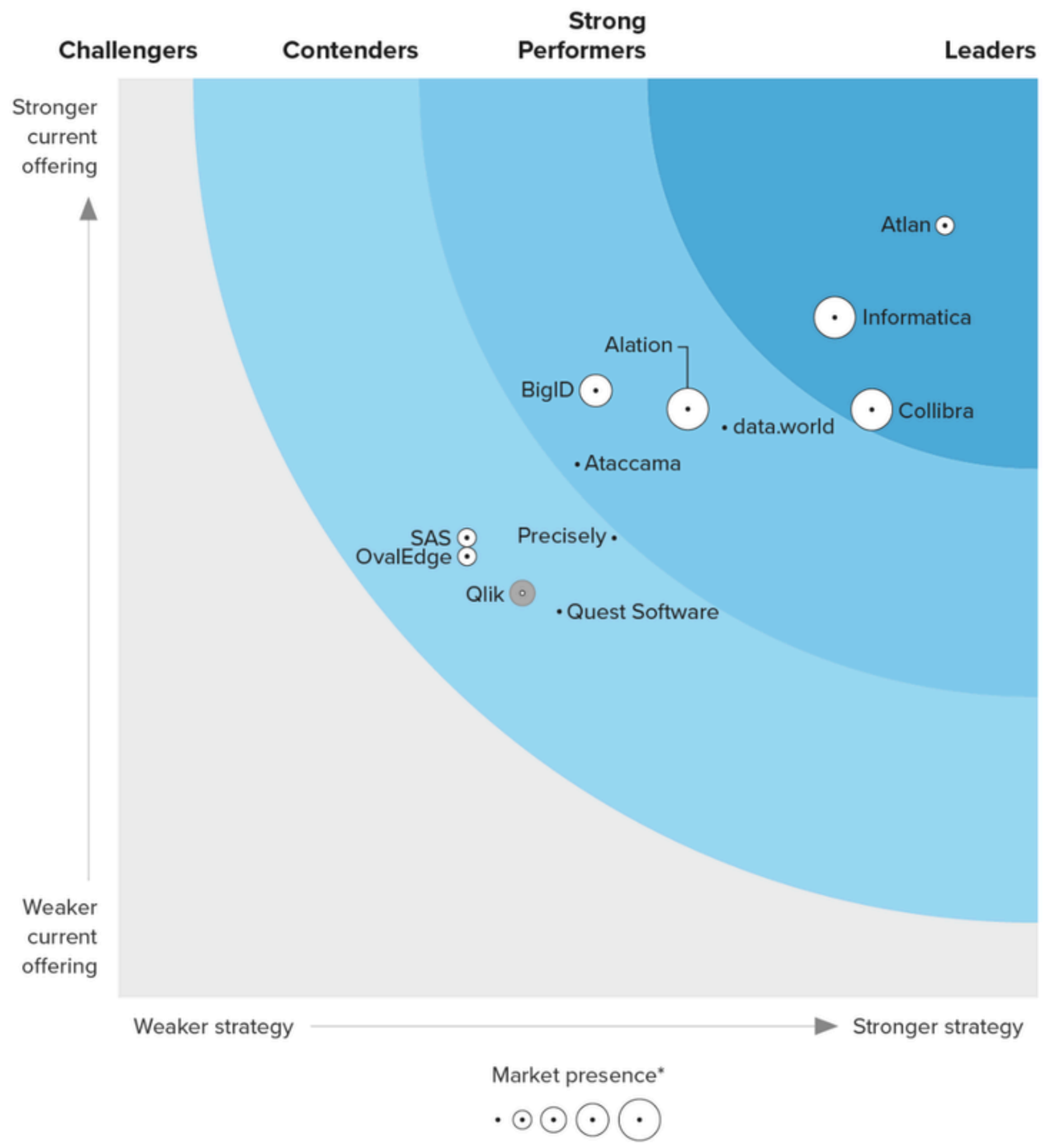
https://www.linkedin.com/in/tomaspeluritis/

https://podcasters.spotify.com/pod/show/duomenu-dede

https://uncledata.substack.com

Data Lineage Market Overview

Forrester Report



As of early 2024, estimates suggested the global data catalog market was valued around $1.5-2 billion

Source: Claude.ai


Alation


Collibra


talend


Open
Metadata


OpenLineage


Apache Atlas

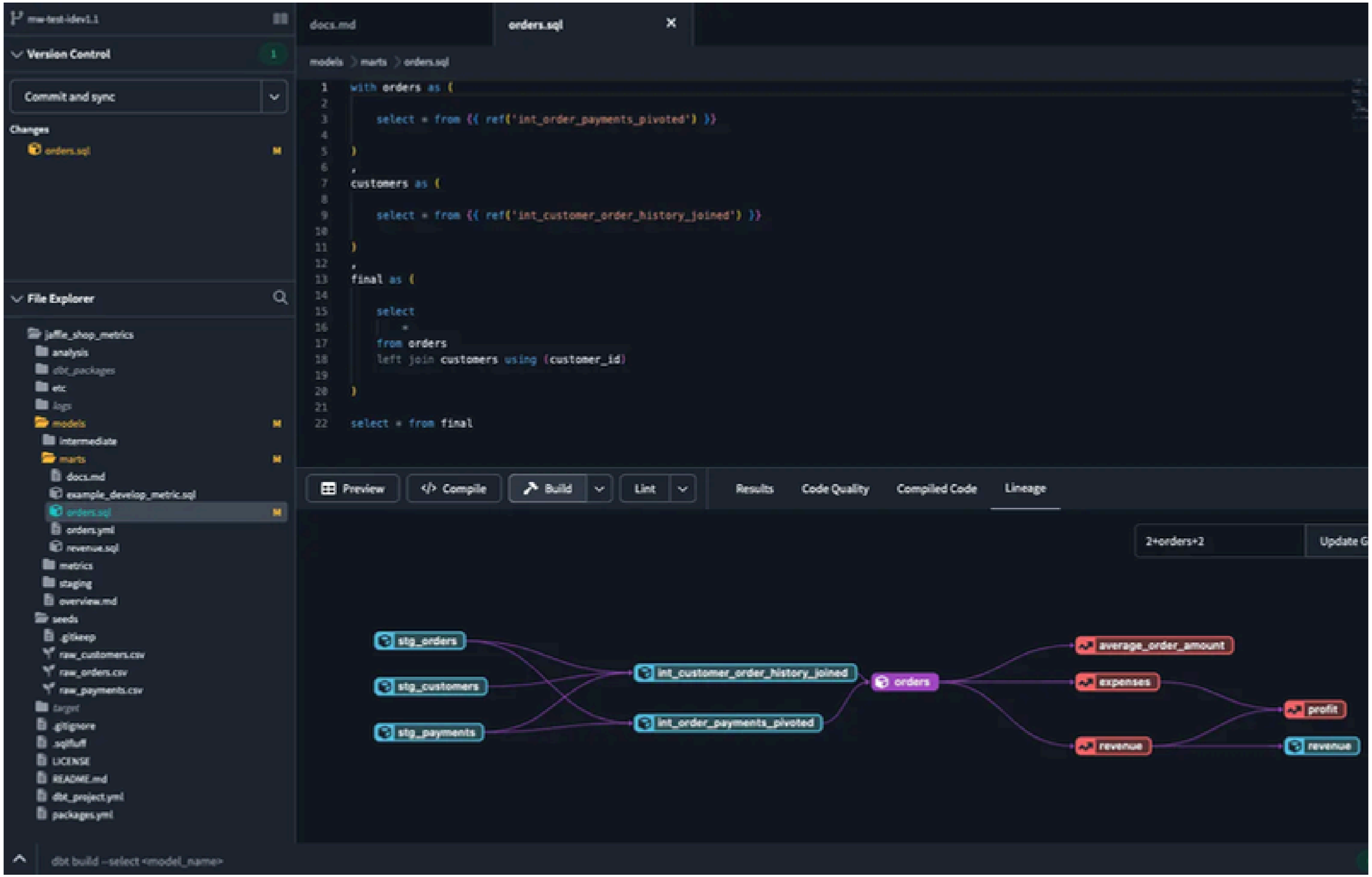

Data
Hub


Acryl Data


atlan

Common Approaches

Just use dbt


Version Control
Commit and sync
Changes
orders.sql
File Explorer
docs.md
orders.sql
models > marts > orders.sql
with orders as (
select * from {{ ref('int_order_payments_pivoted') }}
)
,
customers as (
select * from {{ ref('int_customer_order_history_joined') }}
)
,
final as (
select
*
from orders
left join customers using (customer_id)
)
select * from final
Preview
Compile
Build
Lint
Results
Code Quality
Compiled Code
Lineage
2+orders+2
stg_orders
stg_customers
stg_payments
int_customer_order_history_joined
int_order_payments_pivoted
orders
average_order_amount
expenses
revenue
profit
revenue
dbt build --select <model_name>

Source

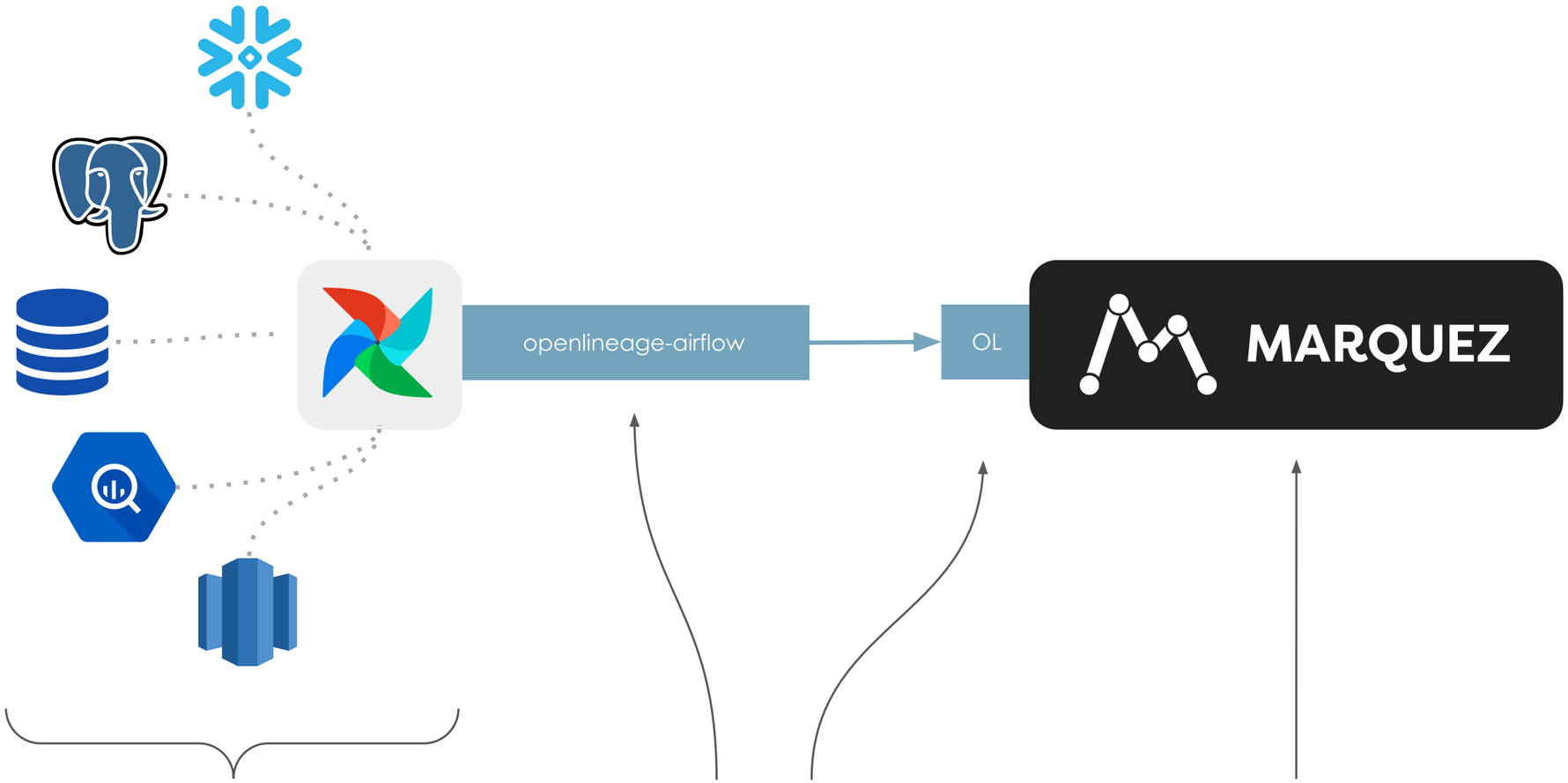
but...


WHERE NON-SQL SOURCES?
imgflip.com

Source


Column level lineage in dbt core
Column-level lineage in dbt cloud ($$$)
imgflip.com

Why not Airflow?



Airflow executes a set of DAGs that operate on various data sources (e.g., Snowflake, Redshift, Bigquery, PostgreSQL)

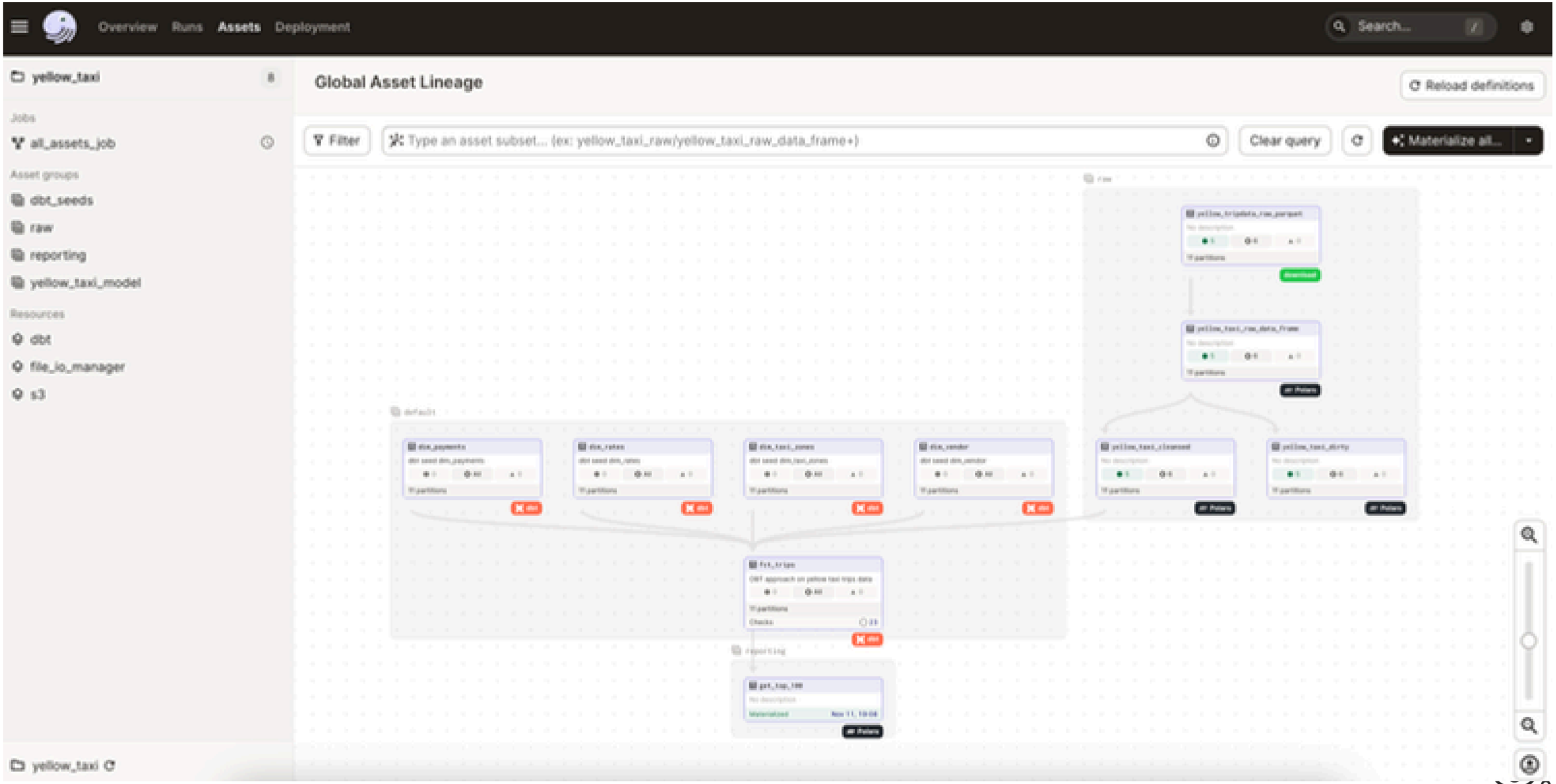
The openlineage-airflow Python module automatically captures metadata and calls the OpenLineage Collection API

Metadata is stored in an OpenLineage-compatible repository like Marquez, where it can be analyzed

but...

- **Not all operators have OpenLineage integrations**
- **Custom Operators require manual OL implementation**
- **SQL Parsing doesn’t work well on custom cases**
- **Column level lineage is not ready yet**
- **Dynamic Tasks + SQL**

Why not Dagster?

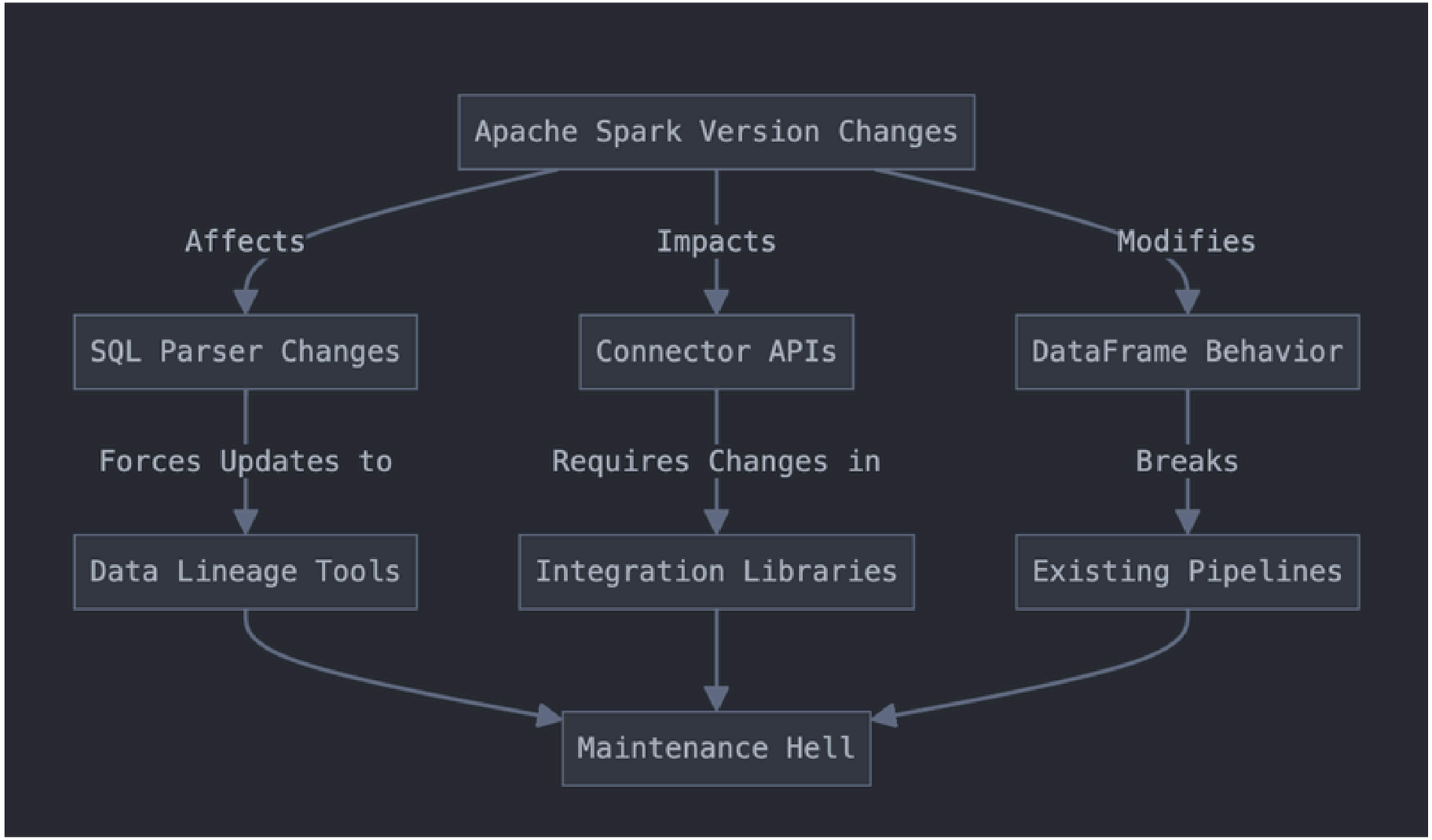

Overview
Runs
Assets
Deployment
Search...
yellow_taxi
Jobs
all_assets_job
Asset groups
dbt_seeds
raw
reporting
yellow_taxi_model
Resources
dbt
file_io_manager
s3
Global Asset Lineage
Reload definitions
Filter
Type an asset subset... (ex: yellow_taxi_raw/yellow_taxi_raw_data_frame+)
Clear query
Materialize all...
default
raw
reporting
dbt
polars
yellow_taxi
yellow_taxi_model

but...

- **Relatively small community**
- **Requires mindset shift**
- **and...**


Column
level lineage
in dagster
Column-level
lineage
in dagster
cloud ($$$)
imgflip.com

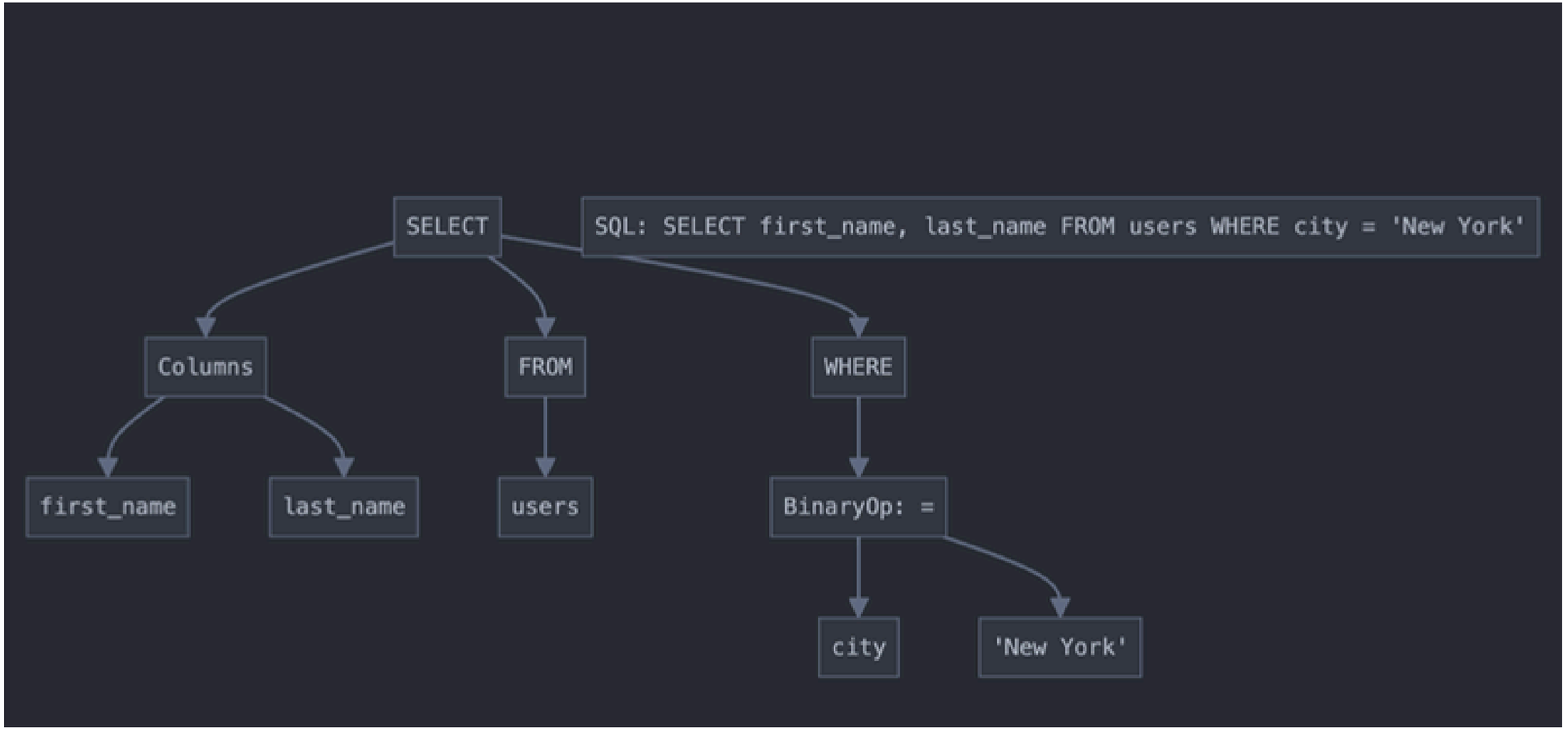
Elephant in the room


Apache Spark Version Changes
Affects
Impacts
Modifies
SQL Parser Changes
Connector APIs
DataFrame Behavior
Forces Updates to
Requires Changes in
Breaks
Data Lineage Tools
Integration Libraries
Existing Pipelines
Maintenance Hell

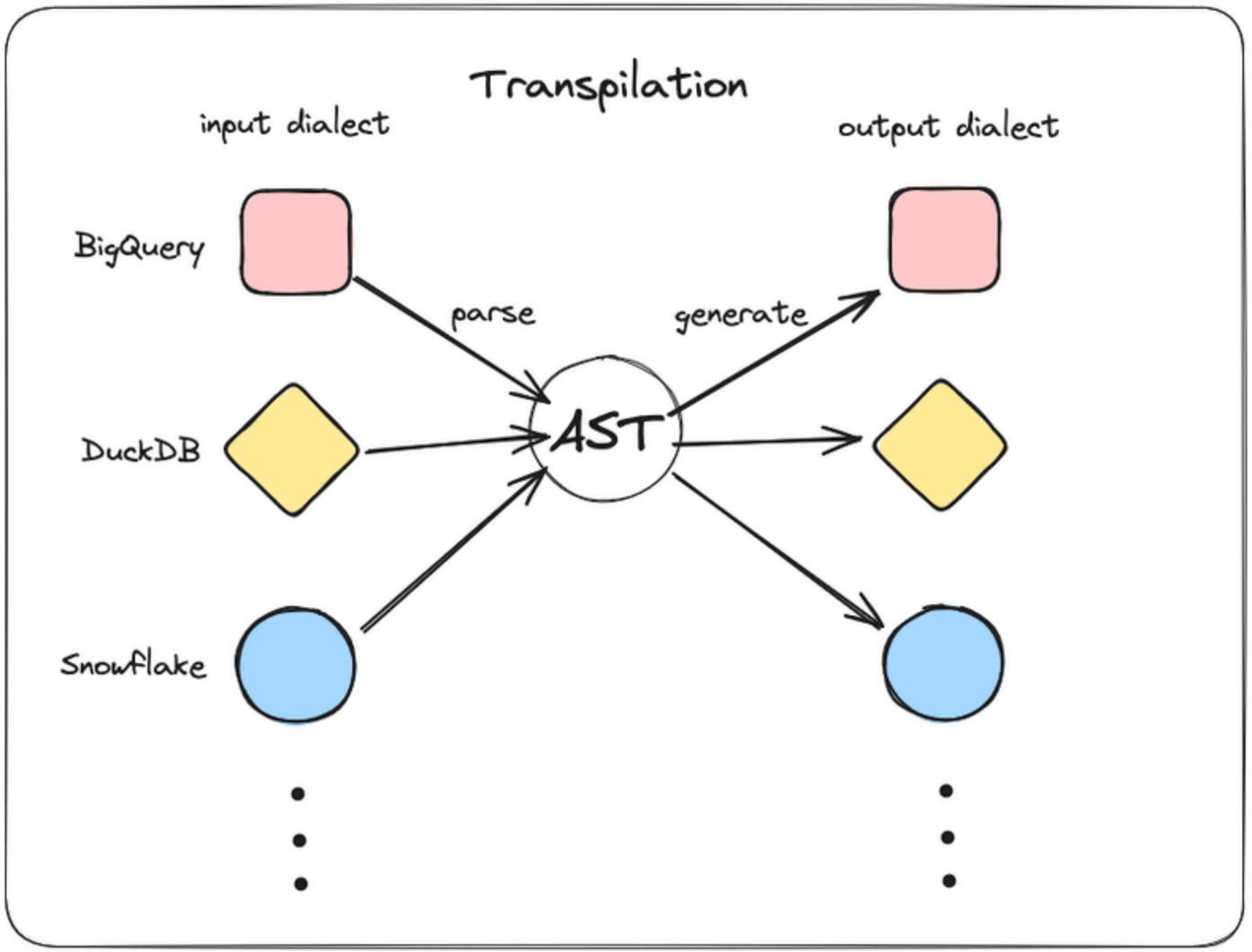
I’m brave ~~stupid~~ enough to DIY

What's SQLGlot?

Abstract Syntax Tree (AST)


SELECT
SQL: SELECT first_name, last_name FROM users WHERE city = 'New York'
Columns
FROM
WHERE
first_name
last_name
users
BinaryOp: =
city
'New York'

Transpilation


Transpilation
input dialect
output dialect
BigQuery
DuckDB
Snowflake
parse
generate
AST

Source

Demo

Questions?