dataroots

a TAIAN company

Internship Project: Open Source Unity Catalog (with dbt +DuckDB)

Unity Catalog

Why does this makes sense?

The Goal

- DuckDb + Unity Catalog Integration with dbt

Why use Unity catalog

- Centralized Governance
- Access management

Why duckDB?

- Fast
- Ease of Setup for analytics

Why dbt?

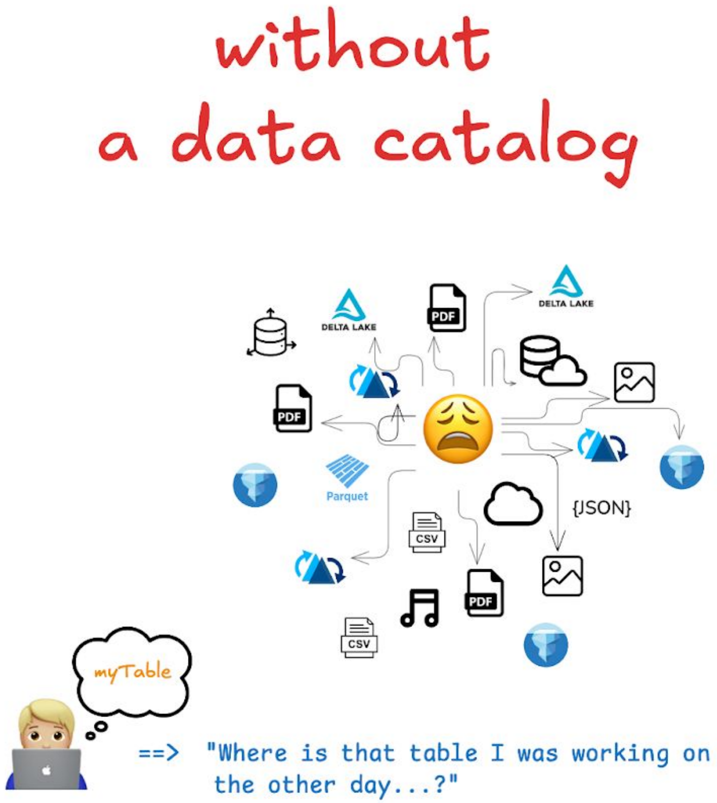
- To automate the the transformation and writing

What is Data Catalog

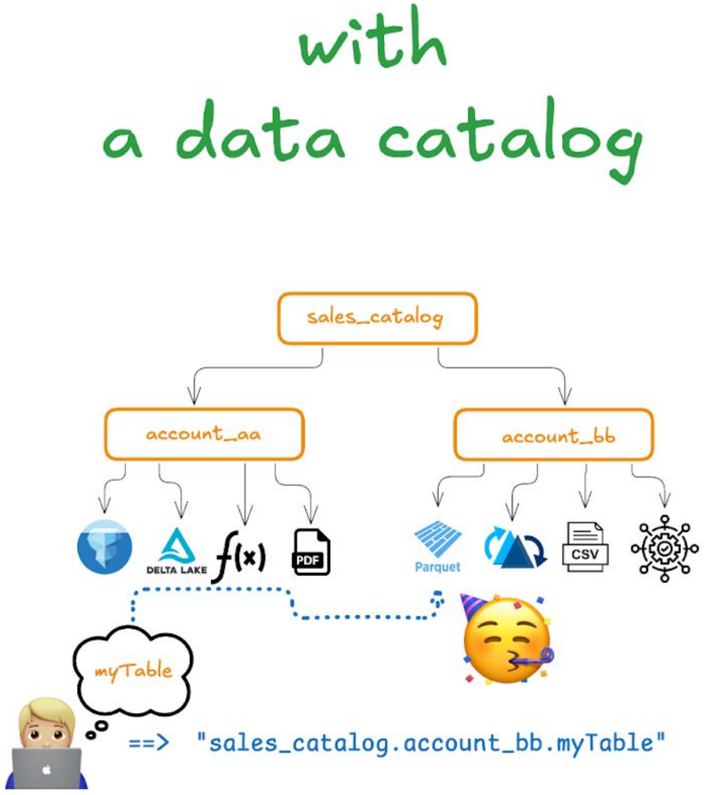
What is Data Catalog ?

Why might you need one ?

- A Metadata layer on top of actual data
- A Centralized inventory of data assets



==> "Where is that table I was working on the other day...?"



==> "sales_catalog.account_bb.myTable"

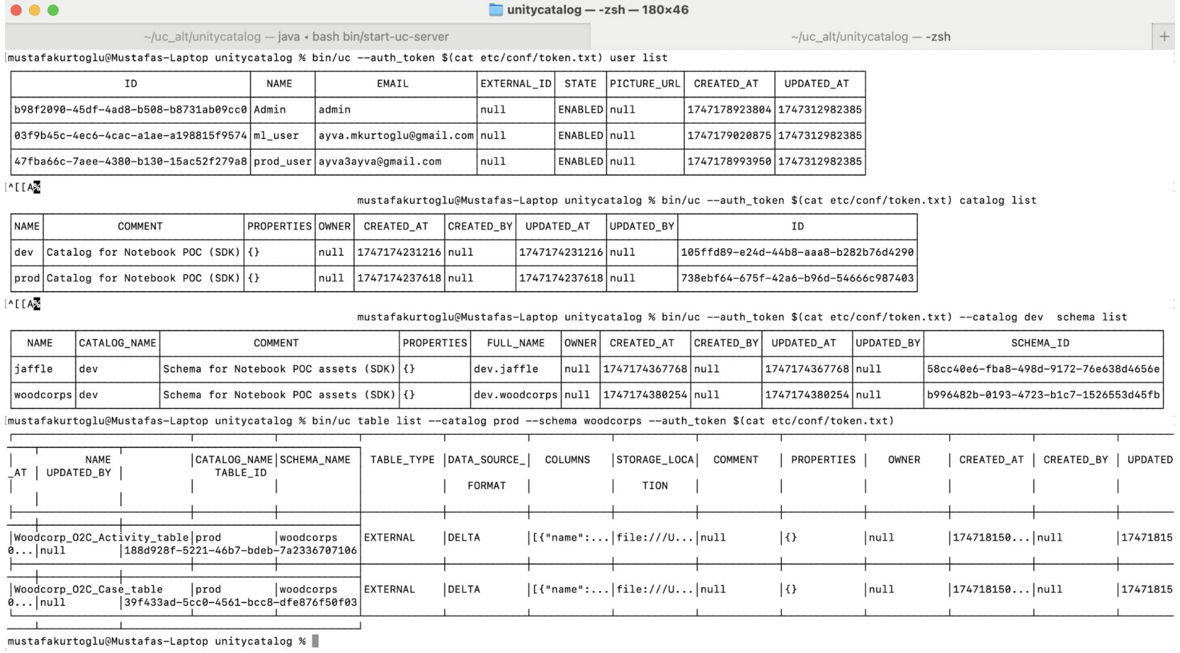
What is Unity Catalog

What is Unity Catalog ?

- **Open source version of Databricks' Unity Catalog. Firstly released in June 2024**

- **Has 2 components, CLI and UI.**

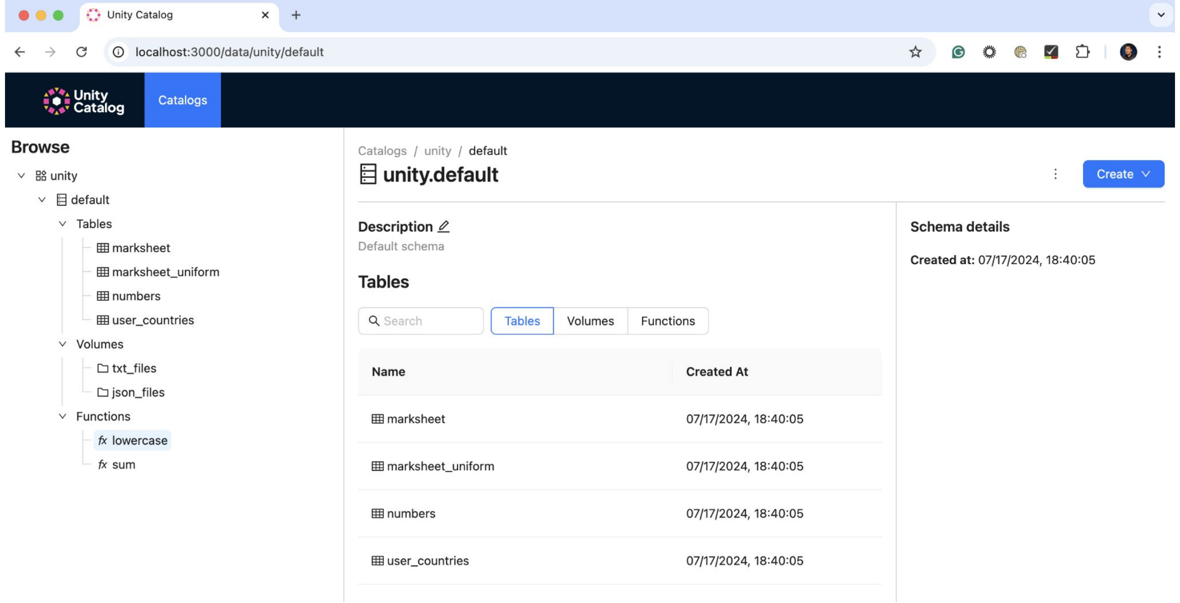
- Command line interface



unitycatalog — -zsh — 180×46
~/uc_alt/unitycatalog — java ◂ bash bin/start-uc-server
~/uc_alt/unitycatalog — -zsh

```
mustafakurtoglu@Mustafas-Laptop unitycatalog % bin/uc --auth_token $(cat etc/conf/token.txt) user list
┌────────────────────────────────────┬─────────┬─────────────────────────┬───────────┬───────┬───────────┬─────────────┬─────────────┐
│                 ID                 │  NAME   │          EMAIL          │EXTERNAL_ID│ STATE │PICTURE_URL│ CREATED_AT  │ UPDATED_AT  │
├────────────────────────────────────┼─────────┼─────────────────────────┼───────────┼───────┼───────────┼─────────────┼─────────────┤
│b98f2090-45df-4ad8-b508-b8731ab09cc0│Admin    │admin                    │null       │ENABLED│null       │1747178923804│1747312982385│
├────────────────────────────────────┼─────────┼─────────────────────────┼───────────┼───────┼───────────┼─────────────┼─────────────┤
│03f9b45c-4ec6-4cac-a1ae-a198815f9574│ml_user  │ayva.mkurtoglu@gmail.com │null       │ENABLED│null       │1747179020875│1747312982385│
├────────────────────────────────────┼─────────┼─────────────────────────┼───────────┼───────┼───────────┼─────────────┼─────────────┤
│47fba66c-7aee-4380-b130-15ac52f279a8│prod_user│ayva3ayva@gmail.com      │null       │ENABLED│null       │1747178993950│1747312982385│
└────────────────────────────────────┴─────────┴─────────────────────────┴───────────┴───────┴───────────┴─────────────┴─────────────┘
^[[A%
                                    mustafakurtoglu@Mustafas-Laptop unitycatalog % bin/uc --auth_token $(cat etc/conf/token.txt) catalog list
┌────┬───────────────────────────┬──────────┬─────┬─────────────┬──────────┬─────────────┬──────────┬────────────────────────────────────┐
│NAME│          COMMENT          │PROPERTIES│OWNER│ CREATED_AT  │CREATED_BY│ UPDATED_AT  │UPDATED_BY│                 ID                 │
├────┼───────────────────────────┼──────────┼─────┼─────────────┼──────────┼─────────────┼──────────┼────────────────────────────────────┤
│dev │Catalog for Notebook POC (SDK)│{}     │null │1747174231216│null      │1747174231216│null      │105ffd89-e24d-44b8-aaa8-b282b76d4290│
├────┼───────────────────────────┼──────────┼─────┼─────────────┼──────────┼─────────────┼──────────┼────────────────────────────────────┤
│prod│Catalog for Notebook POC (SDK)│{}     │null │1747174237618│null      │1747174237618│null      │738ebf64-675f-42a6-b96d-54666c987403│
└────┴───────────────────────────┴──────────┴─────┴─────────────┴──────────┴─────────────┴──────────┴────────────────────────────────────┘
^[[A%
                                    mustafakurtoglu@Mustafas-Laptop unitycatalog % bin/uc --auth_token $(cat etc/conf/token.txt) --catalog dev  schema list
┌─────────┬────────────┬────────────────────────────────┬──────────┬─────────────┬─────┬─────────────┬──────────┬─────────────┬──────────┬────────────────────────────────────┐
│  NAME   │CATALOG_NAME│            COMMENT             │PROPERTIES│  FULL_NAME  │OWNER│ CREATED_AT  │CREATED_BY│ UPDATED_AT  │UPDATED_BY│             SCHEMA_ID              │
├─────────┼────────────┼────────────────────────────────┼──────────┼─────────────┼─────┼─────────────┼──────────┼─────────────┼──────────┼────────────────────────────────────┤
│jaffle   │dev         │Schema for Notebook POC assets (SDK)│{}    │dev.jaffle   │null │1747174367768│null      │1747174367768│null      │58cc40e6-fba8-498d-9172-76e638d4656e│
├─────────┼────────────┼────────────────────────────────┼──────────┼─────────────┼─────┼─────────────┼──────────┼─────────────┼──────────┼────────────────────────────────────┤
│woodcorps│dev         │Schema for Notebook POC assets (SDK)│{}    │dev.woodcorps│null │1747174380254│null      │1747174380254│null      │b996482b-0193-4723-b1c7-1526553d45fb│
└─────────┴────────────┴────────────────────────────────┴──────────┴─────────────┴─────┴─────────────┴──────────┴─────────────┴──────────┴────────────────────────────────────┘
mustafakurtoglu@Mustafas-Laptop unitycatalog % bin/uc table list --catalog prod --schema woodcorps --auth_token $(cat etc/conf/token.txt)
┌────┬────────────┬──────────────────┬────────────┬──────────────┬────────────┬────────────┬────────────┬────────────┬────────────┬────────────┬────────────┬────────────┬────────
│    │      NAME     │                │CATALOG_NAME│SCHEMA_NAME    TABLE_TYPE │DATA_SOURCE_│  COLUMNS   │STORAGE_LOCA│  COMMENT   │ PROPERTIES │   OWNER    │ CREATED_AT │ CREATED_BY │ UPDATED
_AT │ UPDATED_BY │                    TABLE_ID        │
│   │            │                 │            │                           │   FORMAT   │            │    TION    │            │            │            │            │            │
│   │            │
├────┼────────────┼──────────────────┼────────────┼──────────────┼────────────┼────────────┼────────────┼────────────┼────────────┼────────────┼────────────┼────────────┼────────
│Woodcorp_O2C_Activity_table│prod        │woodcorps     EXTERNAL   │DELTA       │[{"name":...│file:///U...│null        │{}          │null        │174718150...│null        │17471815
0...│null        │188d928f-5221-46b7-bdeb-7a2336707106
├────┼────────────┼──────────────────┼────────────┼──────────────┼────────────┼────────────┼────────────┼────────────┼────────────┼────────────┼────────────┼────────────┼────────
│Woodcorp_O2C_Case_table    │prod        │woodcorps     EXTERNAL   │DELTA       │[{"name":...│file:///U...│null        │{}          │null        │174718150...│null        │17471815
0...│null        │39f433ad-5cc0-4561-bcc8-dfe876f50f03
└────┴────────────┴──────────────────┴────────────┴──────────────┘
mustafakurtoglu@Mustafas-Laptop unitycatalog % █
```

- UI



Browse

- unity
  - default
    - Tables
      - marksheet
      - marksheet_uniform
      - numbers
      - user_countries
    - Volumes
      - txt_files
      - json_files
    - Functions
      - lowercase
      - sum

Catalogs / unity / default

unity.default

Description

Default schema

Schema details

Created at: 07/17/2024, 18:40:05

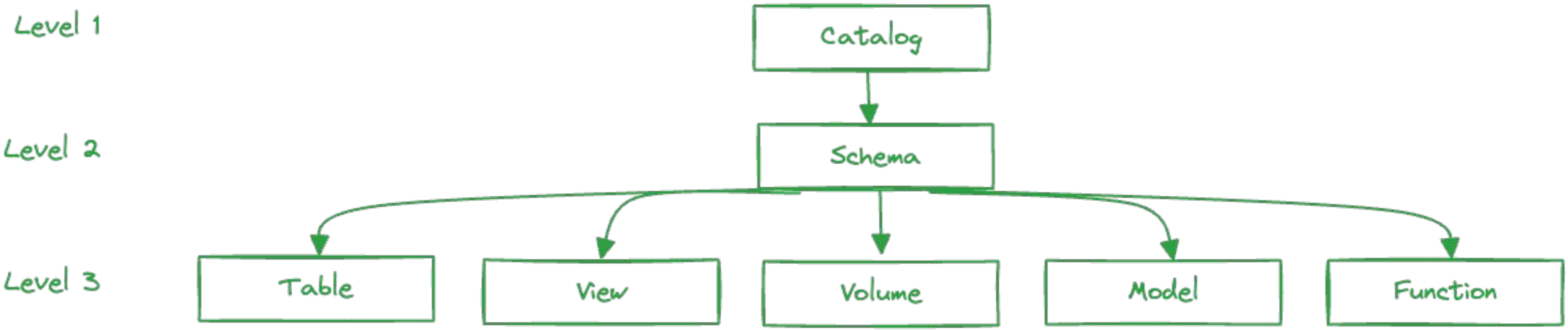
Tables

Name	Created At
marksheet	07/17/2024, 18:40:05
marksheet_uniform	07/17/2024, 18:40:05
numbers	07/17/2024, 18:40:05
user_countries	07/17/2024, 18:40:05

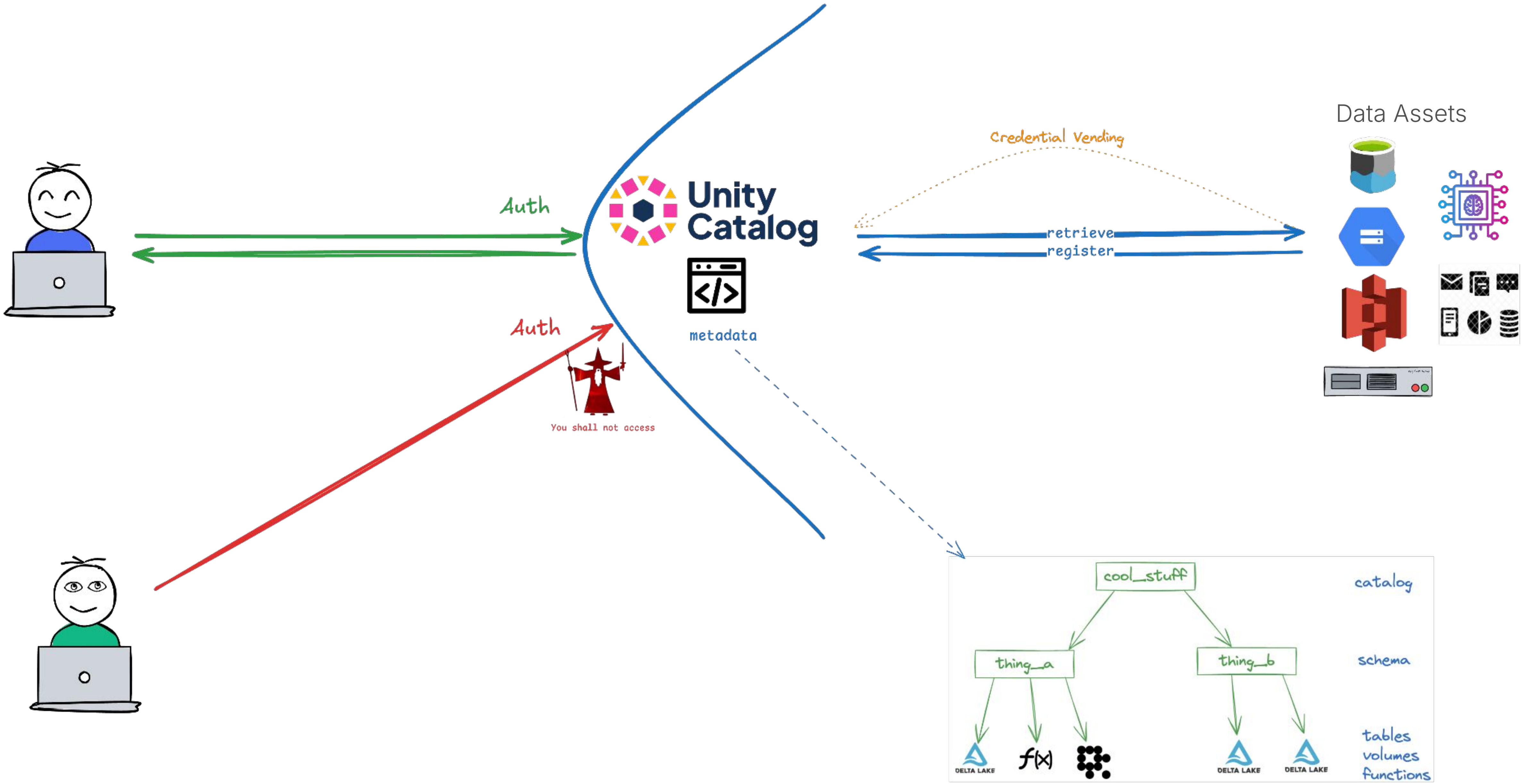
Unity Catalog - Structure

Unity Catalog stores all assets in a 3-level namespace:

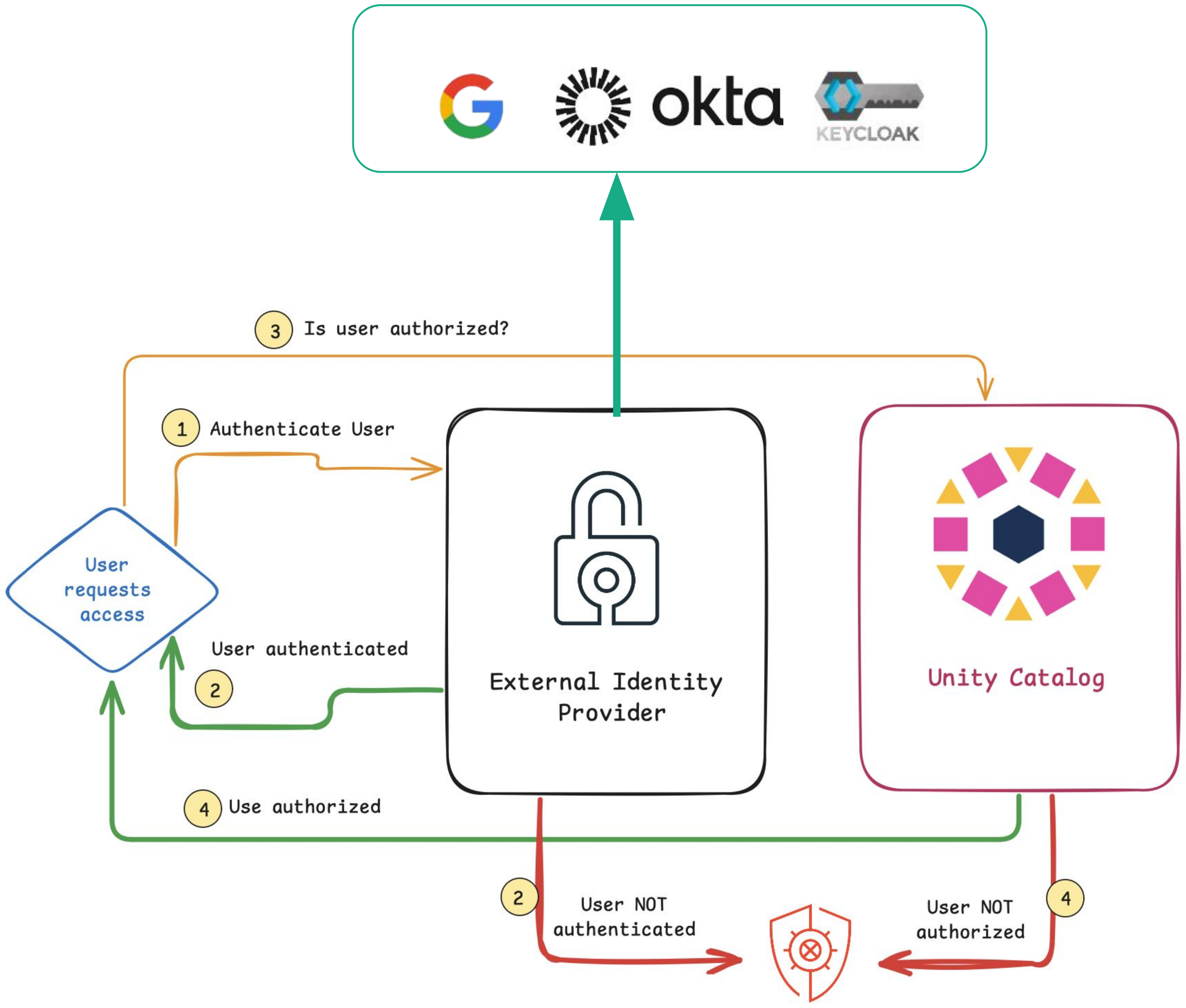
- catalog
- schema
- assets like tables, volumes, functions, etc.



What is Unity Catalog ?



Unity Catalog - Auth



okta

KEYCLOAK

3 Is user authorized?

1 Authenticate User

User requests access

User authenticated

2

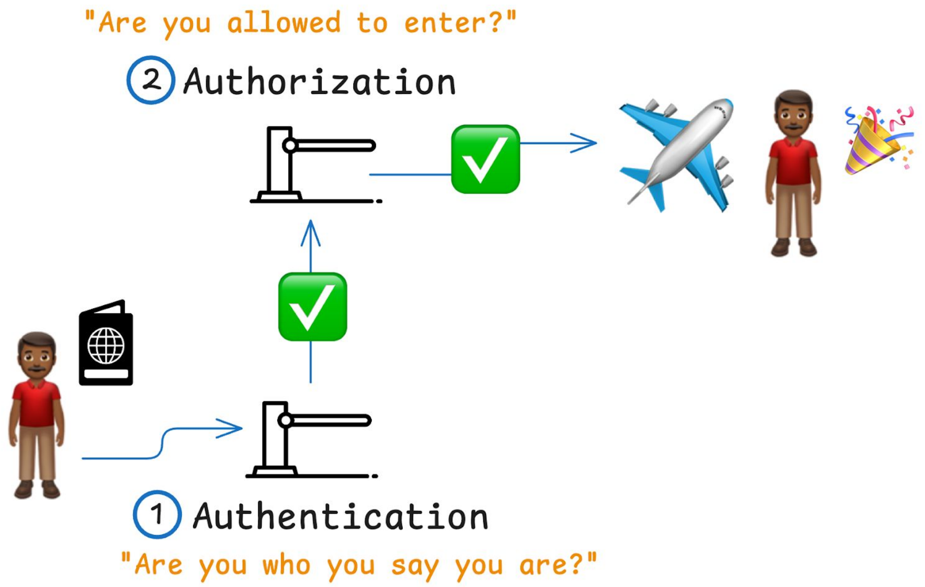
External Identity Provider

Unity Catalog

4 Use authorized

2 User NOT authenticated

4 User NOT authorized



"Are you allowed to enter?"

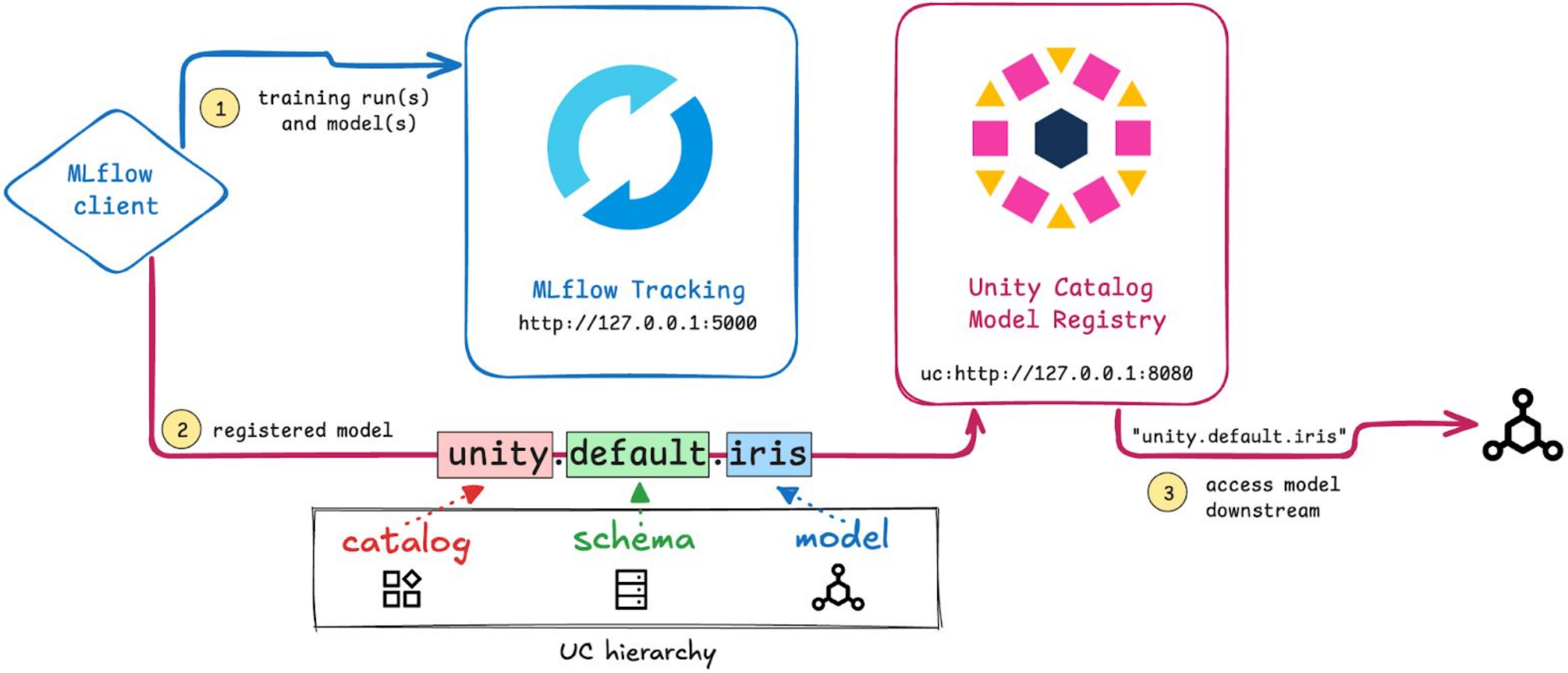
2 Authorization

1 Authentication

"Are you who you say you are?"

Unity Catalog - AI

datar✿ts
a TAIAN company



Unity Catalog Demo

Using CLI

Using Python SDK

Unity Catalog UI

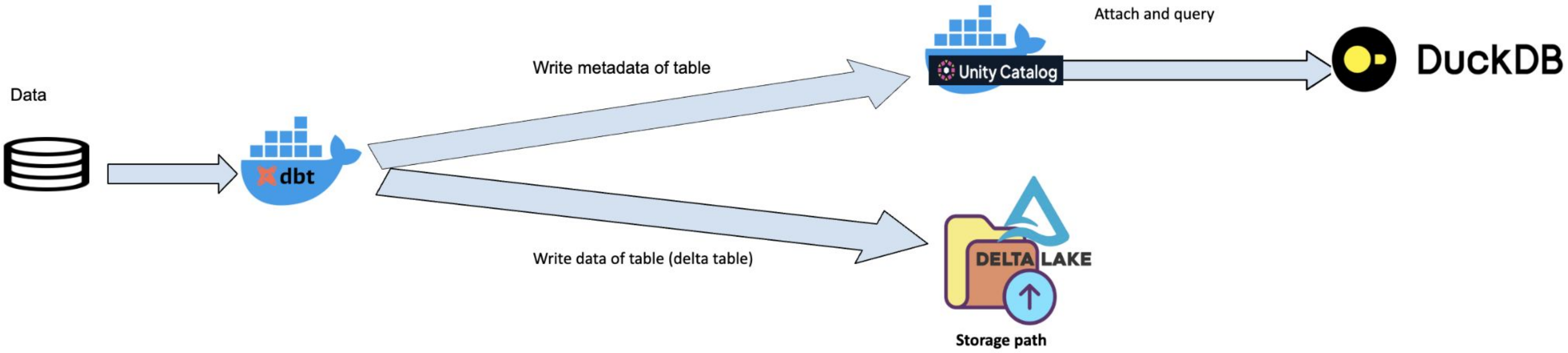
AI

Unity Catalog - Difficulties with Unity catalog

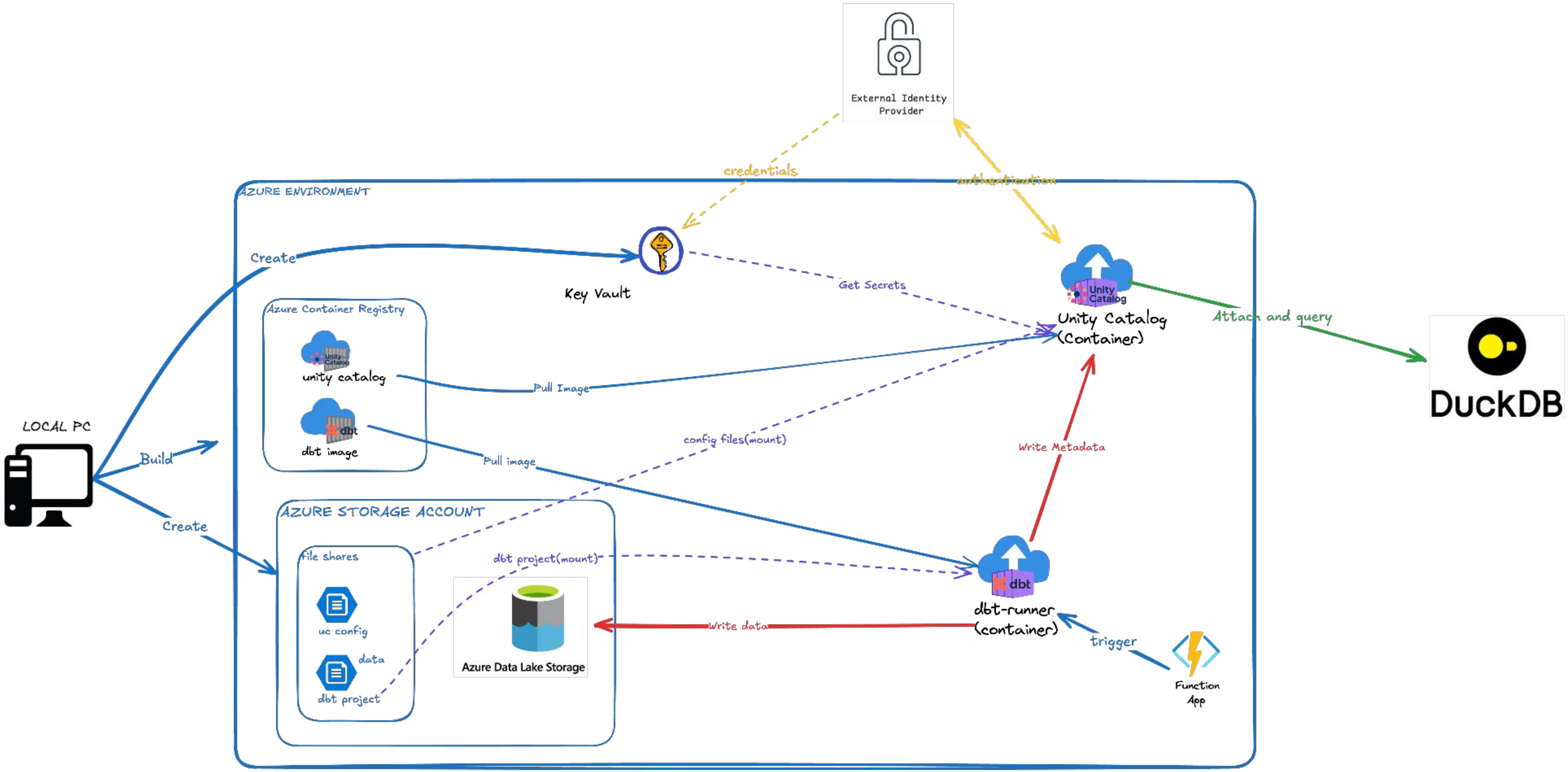
- Writing to unity catalog — dbt
- Querying — DuckDB
- Giving permissions to users — Streamlit

Unity Catalog Project(dbt+azure containers +duckdb)

Unity Catalog Project - Initial Logic (Docker)



Unity Catalog Project - Current Setup (Azure)



Unity Catalog Project Demo

Configurations

dbt

Functions

Azure resources

DuckDB

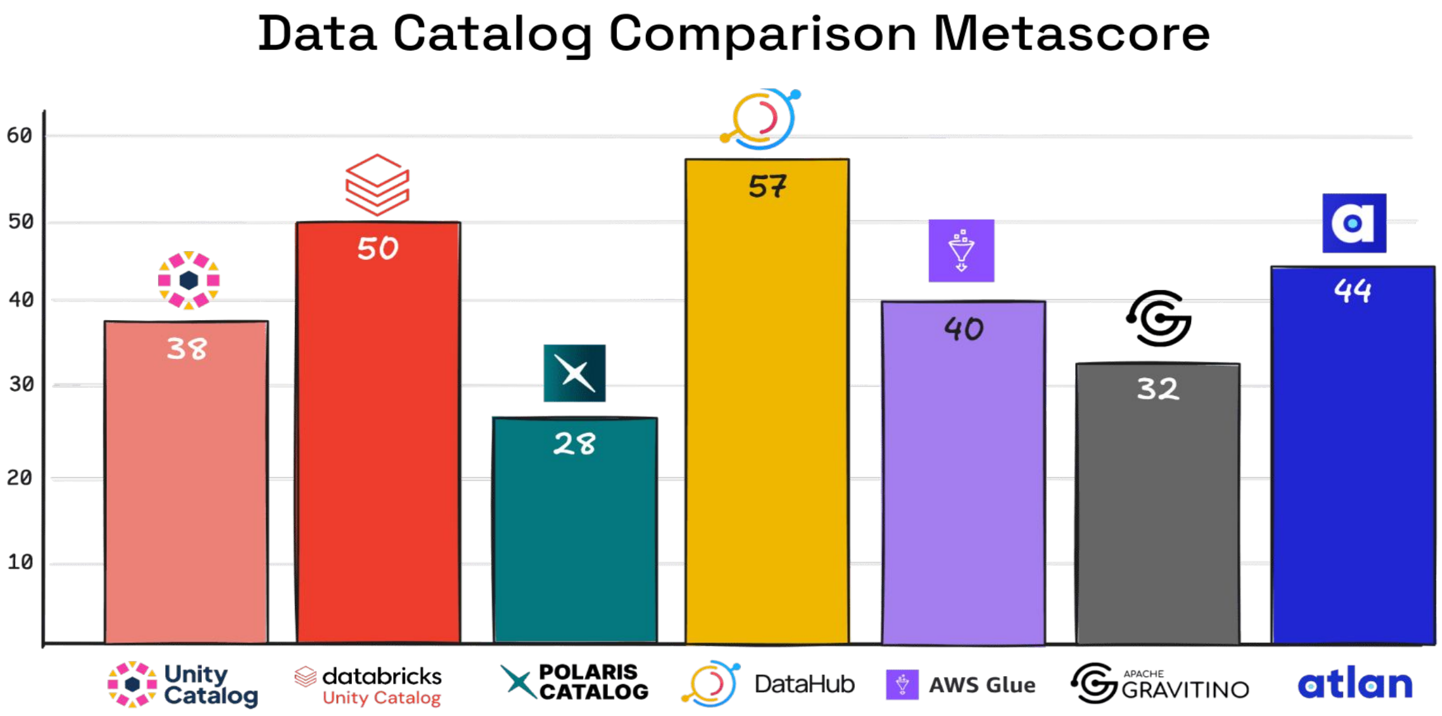
Final Thoughts and limitations

Unity Catalog vs other Catalogs

- **Direct enforcement of access control on the data assets (Gatekeeper)**

- **Unified Governance for Data & AI with Direct Data Lake Control**

- **Structured Namespace for already known from Databricks**



https://www.onehouse.ai/blog/comprehensive-data-catalog-comparison

Unity Catalog - What is missing

THE GOOD

- Unified Governance (tables, volumes, functions, ML models) and access management
- Open Source and free
- Lots of support and integrations(duckdb, spark)
- Lots of room to grow as it is only on 0.2 version

THE BAD

- Limited Data Lineage Capabilities (in OSS):
- No user group permission
- Operational overhead as there is extra layer to manage/debug
- Potential Performance Bottlenecks & Scalability

THE UGLY

- (Currently)Cloud Credential Vending Challenges
- Currently there are some bugs and some things not working as intended
- Maturity Gap Compared to Commercial Offerings:

Additional Use cases

This can be something for you?

Unity Catalog has integrations with :

- Celerdata
- Daft
- DuckDB
- PuppyGraph
- Spark
- SpiceAI
- Trino
- XTable

Also some ai integrations

- LangChain
- LlamaIndex
- OpenAI
- Anthropic
- CrewAI
- AutoGen
- LiteLLM
- Gemini

Thank you and questions