



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

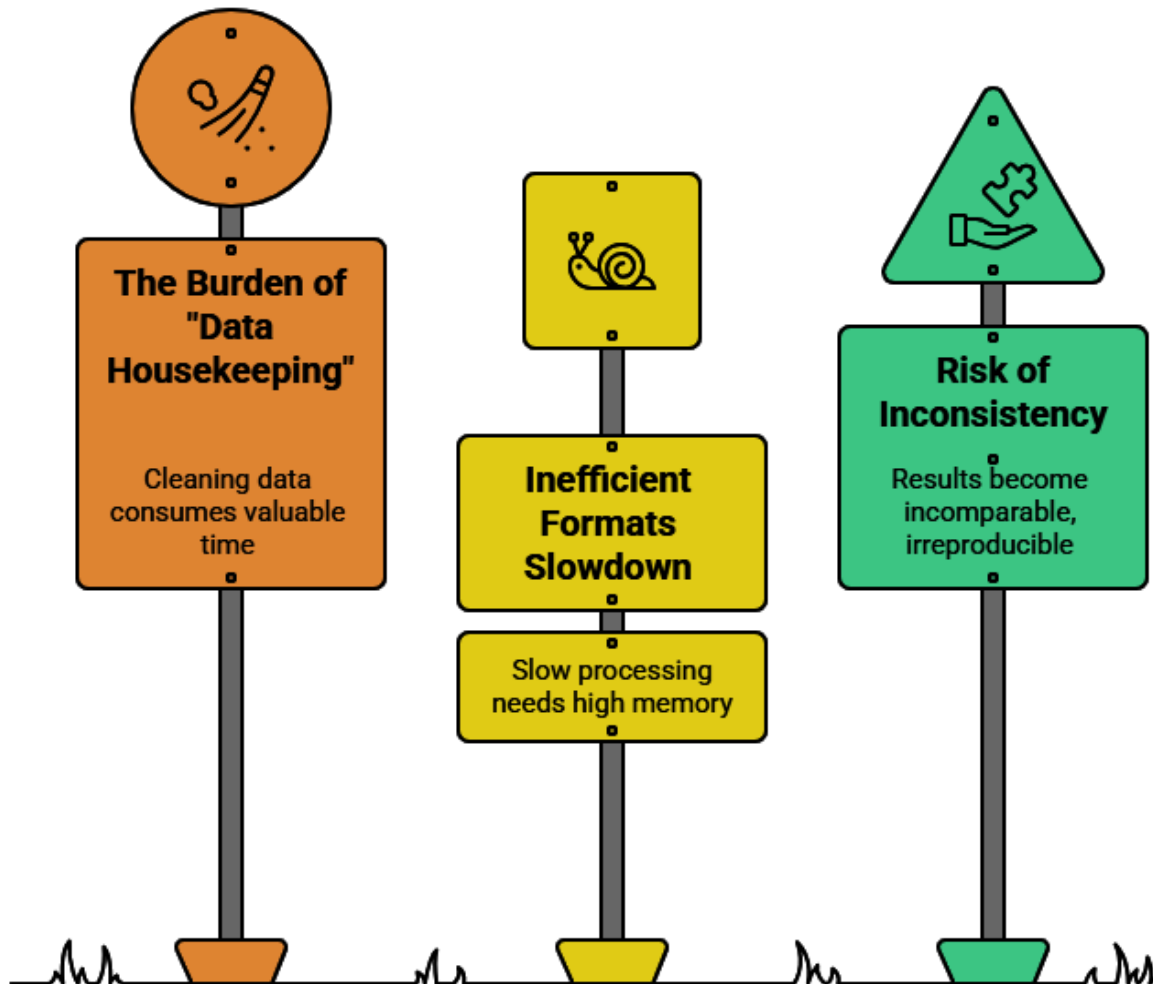
PSDI Data Lakehouse Study

Amali Pawula Hewage

PSDI meeting in Cosener's House,
6-7 October 2025

How the Lakehouse Solves Data Chaos and Empowers Science

The Hidden Cost of "Raw Data"



Analysis, Visualisation, ML



- Effortless Data Access and Discovery
- SQL Superpowers in Your Notebook
- Data is served in the Cloud, no need for local copy

What is a Data Lakehouse?

➤ Data **warehouse** + Data **lake** = Data **lakehouse**

- **Data warehouses** were designed to work really well with structured data.

But...

- Data Warehouse requires you to have your data in a particular order.
- Issue of storing Semi-structured and unstructured data

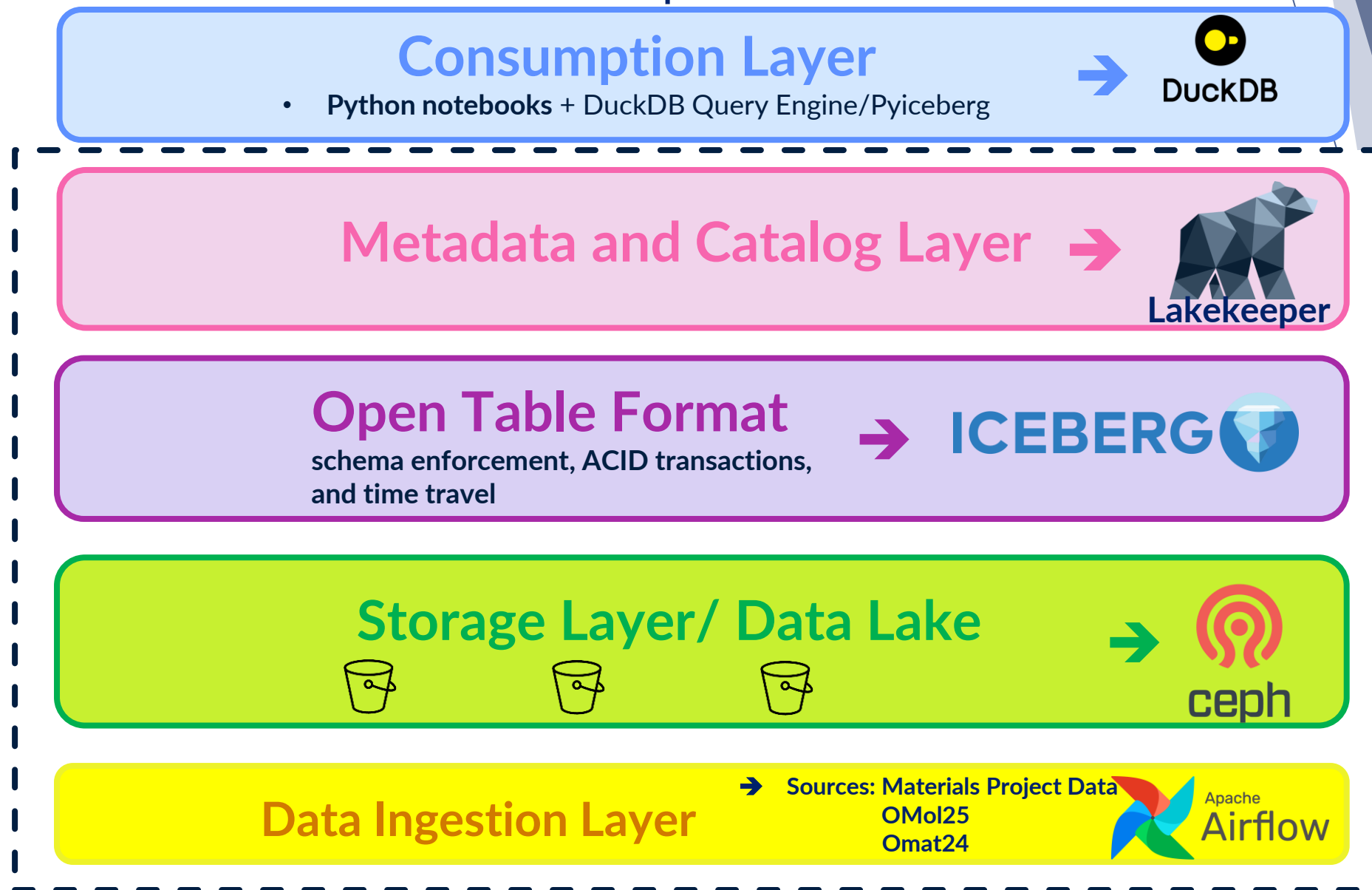
- A **data lake** is big and can hold any raw data, including unstructured data.

But...

- Poor performance for complex queries due to a lack of indexing and optimisation
- No built-in versioning/time-travel capabilities



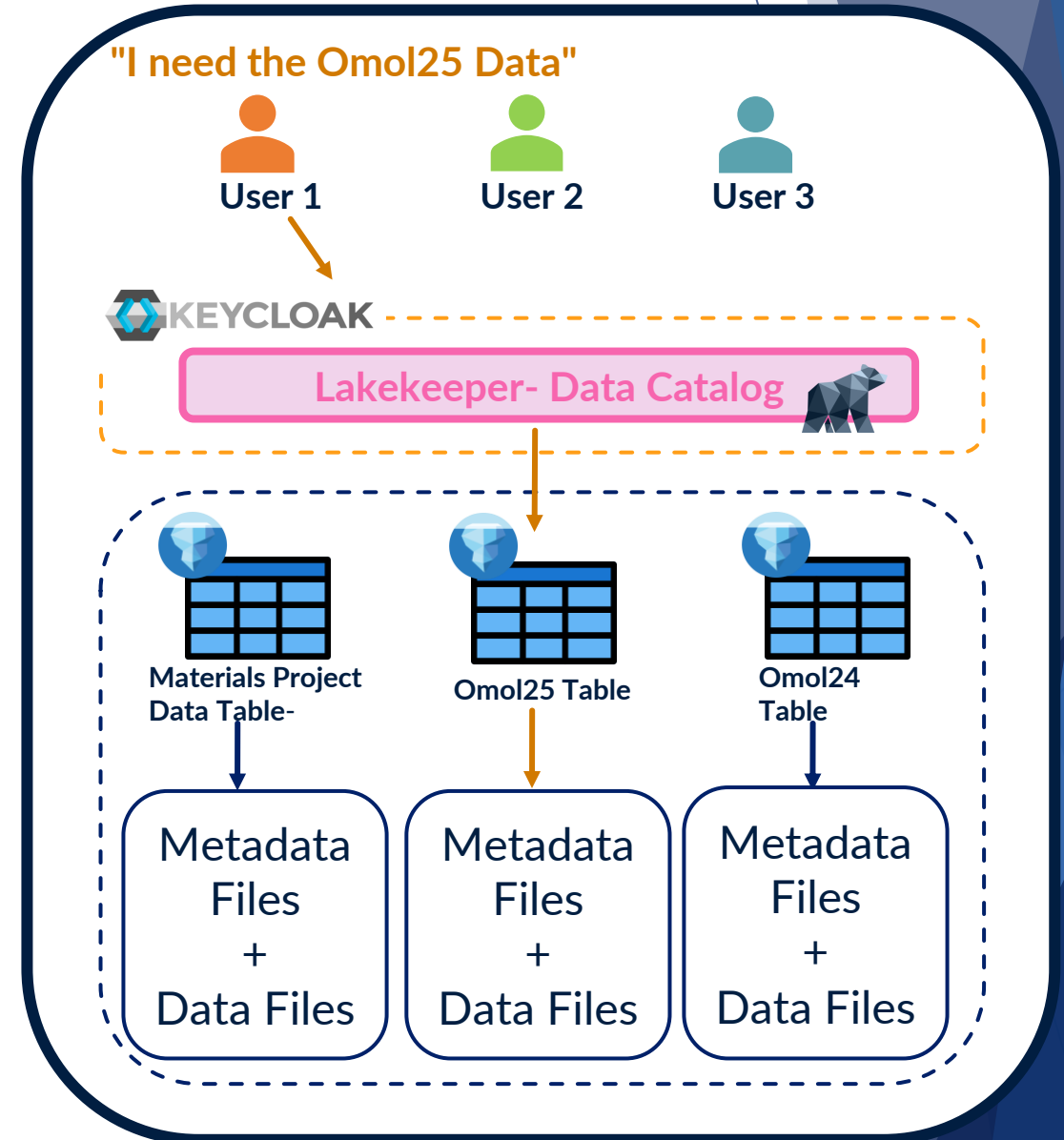
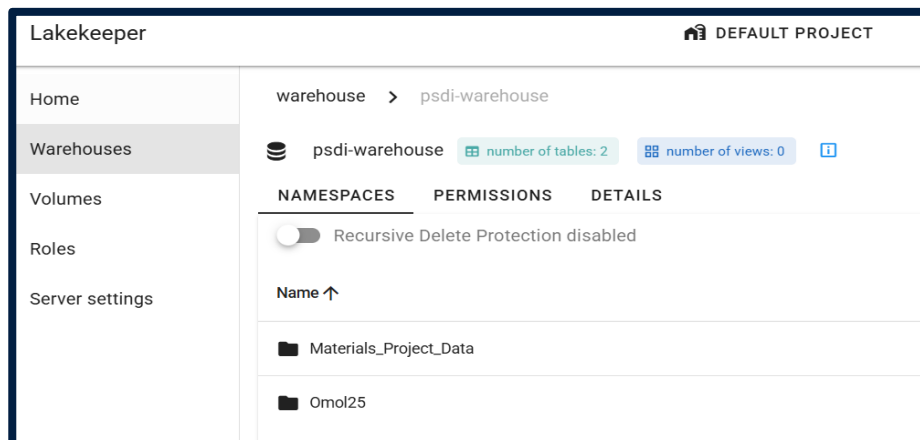
PSDI Data Lakehouse Architecture Components



The Brain of the Lakehouse – Data Catalog

Lakekeeper serves as the catalog for our Iceberg tables

1. **Instant discoverability (like a library catalog)**
2. **Collaboration over datasets with clear schemas**
 - ▶ Everyone points to the same registered table or the dataset with multiple tables
3. **Governance & Trust**
 - ▶ Integration with Keycloak
 - ▶ OpenFGA Authorisation Model



Summary and next steps

- ▶ PSDI data lakehouse could use cost-efficient object storage to serve diverse data collections via uniform interfaces
- ▶ We are on track to having a reasonable lakehouse setup using Open Source components
- ▶ Integration of the lakehouse with client GUIs, computational notebooks or workflow engines should be straightforward
- ▶ In consultation with the project, we can develop a set of examples and training materials for users