



Librarians catalog everything...  
except their OWM data. Let's fix it.

Sierra Overdrive [ data ] → SQL  
lake reports Parquet · DuckDB  
Circ · DuckLake  
"won't this be huge & expensive?"  
a data center vs Parquet > 200 MB  
The whole catalog – on a laptop.

#### HOW IT WORKS

One place that catalogs our own data:

- snapshots at regular intervals – on purpose, not on demand
- metadata – schema, source, capture-time, provenance
- open formats (Parquet) – no proprietary readers
- time travel – query the collection as it was on any date

#### WHAT A DATA LAKE IS

Pull a report today? Easy. Compare it to last month? Hard.

- snapshots are ad-hoc – only when someone asks
- no record of when or how it was captured
- updates overwrite the past – the old version is just gone

Our data is always moving.

#### THE PROBLEM

#### WHY IT MATTERS HERE

It's a **security win**, not just analytics:

- **retention** – expire patron data on a schedule, not by accident
- **access control** – the lake is the boundary; no patron data on personal laptops
- **analysis runs at the engine** – reports come out, raw data stays put

#### IT COMPOUNDS

Same pattern, every new source:

- circulation transactions
- COUNTER 5 (e-resource usage)
- door counts, holds, ...

one pipeline at a time  
the pattern just scales



scan it > see the talk

[rayvoelker.github.io/2026-05/issue\\_2\\_our-data-our-lake/](https://rayvoelker.github.io/2026-05/issue_2_our-data-our-lake/) · v1.0.0

- - - dub & pass it on - - -

OUR DATA  
OUR LAKE



a data lake for the library  
own it · snapshot it · keep the history  
issue #2 · may '26