

+-----+
| O U R D A T A , O U R L A K E |
| a data lake for the library |
+-----+

own it // snapshot it // keep the history // issue #2 // may '26

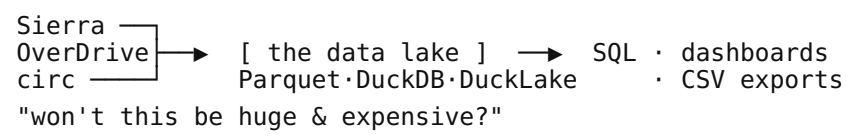
Pull a report today? Easy. Compare it to last month, or last year? *Hard* – the underlying data has already moved on. Librarians catalog everything... except their own data. **A data lake fixes that.**

WHAT A DATA LAKE IS

One place that catalogs **our own data**, deliberately and over time:

- **Snapshots** at regular intervals – captured on purpose, not on demand
- **Metadata** – schema, source, capture-time, provenance (a real catalog record)
- **Open formats** (Parquet) – no proprietary readers, nothing to rent back
- **Time travel** – query the collection *as it was* on any past date

HOW IT WORKS :: it runs on a laptop



a data center 
vs
Parquet ■ < 200 MB

Free, open tools – the **entire CHPL catalog fits in under 200 MB**: columnar, compressed, no data center required.

WHY IT MATTERS HERE :: a security posture, not just analytics

- **Retention** – expire patron data on a schedule, not by accident
 - **Access control** – the lake is the boundary; no patron data on laptops
 - **Analysis at the engine** – reports come out, the raw data stays put
- Our data belongs to us. **The capability to keep, version, and protect it is the story.**

IT COMPOUNDS

One pipeline at a time, the pattern just scales: circulation transactions, COUNTER 5 e-resource usage, door counts, holds. Start small; add as you go.



scan it ▶ see the talk
rayvoelker.github.io/2026-05/issue_2_our-data-our-lake/
v1.0.0 · 2026-05

- - - dub & pass it on - - -