

Cloud-Native Graphical and Programmatic Cross-Database Access to Experiment Metadata with Trino, Superset and Ibis

Authors

Eli Mizrachi <emiz@slac.stanford.edu>, SLAC National Accelerator Laboratory, The LZ Collaboration, 0000-0001-6047-3588

Keywords

data engineering, trino, superset, ibis, helm, SQL, dataframes, python, data analysis, data lineage, metadata, data management, pipelines, ETL

Abstract

LUX-ZEPLIN (LZ) is the world's most sensitive dark matter direct-detection experiment [1], acquiring over a petabyte of data per year [2]. LZ also retains several terabytes of user-facing metadata describing detector conditions and data-processing environments needed for reproducible data selection and processing. These metadata live in six heterogeneous databases (SQL and NoSQL) and have historically been accessed through seven separate programmatic and graphical interfaces that connect to unique subsets of these databases.

This fragmented nature has led to major problems for both users and maintainers. First, maintenance overhead grew disproportionately: for example, our custom python-based GUI for browsing databases grew to thousands of lines of non-reusable code, requiring near-daily manual restarts to recover from silent failures reported to us by users. Second, our programmatic python interfaces lacked expressiveness, pushing users toward ad-hoc scripts to filter/reshape results and making cross-database joins unnecessarily tedious.

More recently, the Genesis Mission and the advent of the American Science Cloud [3] from the US Department of Energy has indicated strong institutional interest in implementing open-source tools and services from the data engineering ecosystem in order to enhance compatibility with AI-based workflows. Recognizing these factors and the aforementioned problems, we consolidated access around three complimentary tools: Trino [4], a distributed SQL query engine which provides a single SQL facade across databases; Superset [5], a graphical dashboard builder which operates directly on Trino; and Ibis [6], a library to create composable SQL queries with dataframe-like python API.

In this talk I will detail the benefits we recouped from deploying these tools, along with practical challenges encountered along the way, such as reconciling Trino's SQL interface with schemaless MongoDB sources. Topics covered include a showcase of how we were able to immediately deprecate one database browser with Superset and Trino, consolidating a key dashboard and cutting worst-case load times from 12s to 3–4s. We also reduced a previously expensive categorical lookup from 40s to 100ms by materializing a pre-aggregated table with Ibis. I will also discuss using Ibis to improve transparency and expressiveness over our previous python interfaces, while keeping a smaller code footprint. Finally, I will touch on the advantages of our shift to a GitOps workflow. Helm cut deployment time for these services from multiple hours of manual configuration via a GUI to a single reproducible command, while Ibis-defined data flows now refresh derived datasets via scheduled CI/CD pipelines instead of requiring manual input.

References

- [1] D. S. Akerib *et al.*, “Searches for Light Dark Matter and Evidence of Coherent Elastic Neutrino-Nucleus Scattering of Solar Neutrinos with the LUX-ZEPLIN (LZ) Experiment,” Dec. 23, 2025, *arXiv*: arXiv:2512.08065. doi: 10.48550/arXiv.2512.08065.
- [2] M. Arthurs *et al.*, “The Real-Time Data Workflow of The LUX-ZEPLIN Dark Matter Experiment at NERSC,” *EPJ Web Conf.*, vol. 337, p. 01232, 2025, doi: 10.1051/epjconf/202533701232.
- [3] “Introduction to the American Science Cloud (AmSC): Part 2,” Jan. 20, 2026. [Online]. Available: https://american-science-cloud.github.io/amsc-site/events/AmSC_January_Webinar_Part2_Slides.pdf
- [4] *trinodb/trino*. (May 14, 2025). Java. Trino. Accessed: May 14, 2025. [Online]. Available: <https://github.com/trinodb/trino>
- [5] *apache/superset*. (May 15, 2025). The Apache Software Foundation. Accessed: May 14, 2025. [Online]. Available: <https://github.com/apache/superset>
- [6] The Ibis Project Authors, *The Ibis Project*. (Aug. 12, 2025). Python. Accessed: Aug. 12, 2025. [Online]. Available: <https://github.com/ibis-project/ibis>

Connection to Mission, Goals, & Interests of US-RSE Community

Data management is a complex topic, and the rise of AI/ML has accentuated the need for engineered solutions and the employment of best practices to handle large volumes of data and/or data from disparate sources. I believe this work will be of interest to RSEs because it demonstrates an easy-to-generalize application of open-source tools that drastically simplify and improve access to data spread across multiple databases (of varying dialects!) in an academic context. Beyond the specific tools, I hope to leave attendees with these takeaways: a framework for evaluating and selecting BI/query tooling for heterogeneous data, patterns for keeping query logic reusable across dashboards and scripts, and lessons learned from adopting a GitOps deployment model for research infrastructure. Finally, growing institutional interest in AI-ready data infrastructure, e.g. the Genesis Mission and American Science Cloud, suggests that investments like these will pay continued dividends as AI-based workflows become more common in scientific computing.