Contribution ID: **25** Type: **Poster**

Towards a Unified Lakehouse Platform for PSDI

Tuesday 4 November 2025 13:40 (5 minutes)

PSDI is the UK's nationally funded programme that provides tools and services to help researchers in the physical sciences find, share, and process data, with the explicit aim of accelerating scientific discovery and innovation. In PSDI, we work with diverse data from various sources. One of the key challenges we face is managing big data while maintaining flexibility in handling both raw and complex data in low-cost storage, and addressing issues related to data governance, performance, and consistency. To truly empower the scientific community, this data must be usable for both analytics and cutting-edge AI/ML applications.

To tackle this, we will design and build a 'data lakehouse'on low-cost object storage. This architecture combines the flexibility of a data lake with the transactional consistency and query performance of a data warehouse. Raw datasets from different data sources will be ingested into object storage and transformed into a common, open format like Apache Parquet, enabling efficient analytics. These datasets will then be registered as Apache Iceberg tables in a metadata catalog (e.g., Lakekeeper or Apache Polaris) to manage schema and ensure consistency. By providing a unified, governed platform with a powerful query engine (e.g., Trino, DuckDB, or Spark), this lakehouse will make diverse data more Findable, Accessible, Interoperable, and Reusable (FAIR). Ultimately, this work will make it easier for the scientific community to exploit this data for new insights and discoveries.

References:

1. PSDI
2. Guide to Architecting-an-iceberg-lakehouse

Confirm eligibility

Author: PAWULA HEWAGE, Amali (UKRI - STFC)

Co-authors: Dr BELOZEROV, Alexander; Dr UNDERWOOD, Tom; Dr BUNAKOV, Vasily

Presenter: PAWULA HEWAGE, Amali (UKRI - STFC)

Session Classification: Poster Session

Track Classification: Poster