Analyzing and Visualizing the Migration of Scientific Data at NSLS-II

Nawal Sarwar

Abstract

At National Synchrotron Light Source II (NSLS-II), efficient access to data is essential for supporting high-capacity data-driven science. The facility's data infrastructure is built on the Bluesky data collection framework and the Tiled data management system, which enables users to retrieve, explore, and analyze scientific datasets across nearly all beamlines and several partner institutions. Tiled was originally deployed with a NoSQL backend, but evolving database technology and a clearer understanding of access patterns at NSLS-II have motivated a transition to a SQL-based storage system.

This project evaluates the integrity of the data migration, which encompasses nearly 10 petabytes of scientific data. We developed validation scripts that compare query outputs from both database systems, log success and failure rates, and produce visual summaries of migration activity. Through this work, I gained experience in data validation, large-scale system monitoring, and scientific database management-skills essential for maintaining and evolving complex research infrastructure in modern experimental facilities.

Introduction

The National Synchrotron Light Source II (NSLS-II) is a premier research facility that produces enormous volumes of scientific data across its many beamlines. Our data management system, Tiled, has powered seamless access to this information using a NoSQL backend for the past decade. However, advancements in database technology and our deeper insight into how users work with the data have led us to transition from NoSQL to a faster, more robust SQL backend.

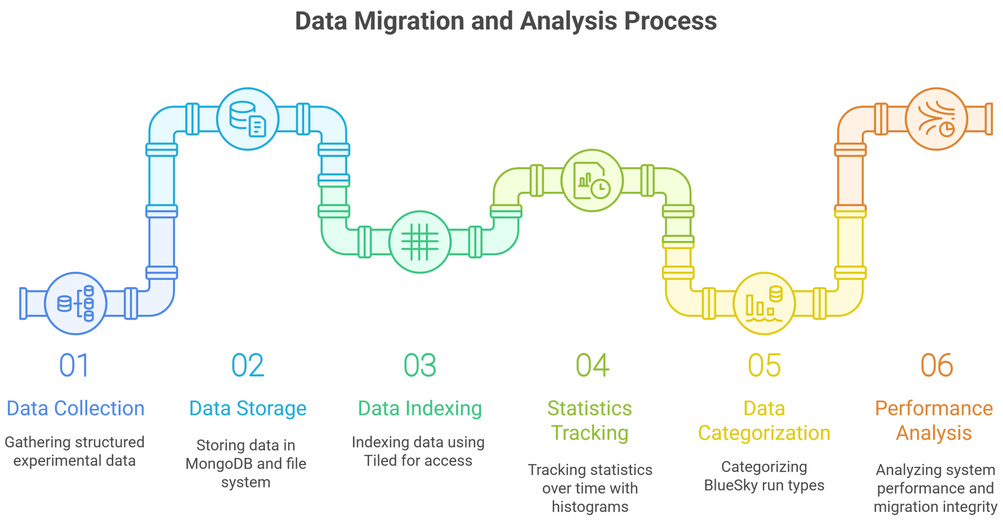
To test this migration, we've set up the new SQL-backed Tiled on the Hard X-ray Nanoprobe (HXN) beamline-a high-resolution X-ray station used for detailed imaging and spectroscopy. HXN generates large, complex datasets, making it the perfect proving ground for our new system. Our goal is to verify that every piece of data (nearly 10 PB in total) is preserved exactly as before. We're building automated scripts to compare outputs between the old and new databases, gather statistics on any differences, and visualize the results. A successful pilot on HXN will clear the way for rolling out the SQL backend across the entire facility, delivering faster access, simpler administration, and exciting new features like contextual search.

Methodology

Automate Extraction & Comparison with Python: Use a Python script to call both the MongoDB and SQL Tiled APIs, retrieve matching records, and flag any discrepancies for further investigation

Track Statistics Over Time: Generate histograms at regular intervals to visualize how data characteristics and migration outcomes evolve, highlighting any trends or delays.

Categorize BlueSky Run Types in MongoDB: Query the MongoDB metadata to tally each run type and compute daily totals, providing insight into dataset diversity and baseline load.



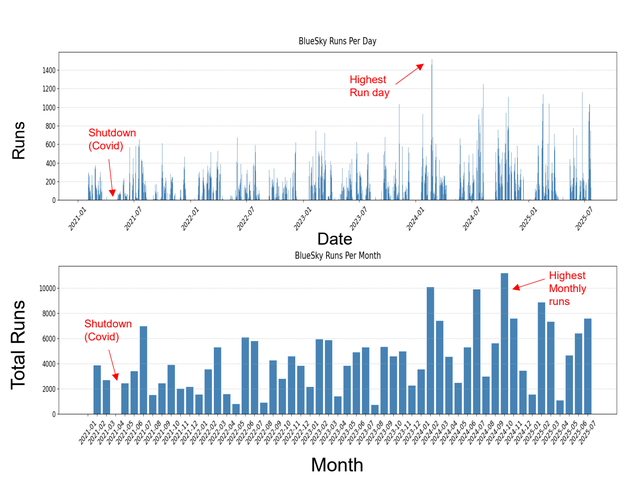
Future Plans

Extend the Time Horizon
Retrieve and analyze daily-run data back to 2016 to establish a longer baseline for trends, seasonality, and baseline load-giving context to the post-COVID ramp and helping distinguish long-term shifts from short-term effects.

Incorporate Full Data Streams
Move beyond metadata counts by examining raw data attributes (file sizes, acquisition durations, beamline parameters) to understand how data volume and complexity have evolved, and to spot any performance or storage bottlenecks.

Assess Migration Coverage
Parse migration logs and summary documents to quantify exactly how much data (in bytes, run counts, and experiment types) has been successfully migrated versus any hold-outs or failures, and chart those metrics over time to ensure complete coverage.

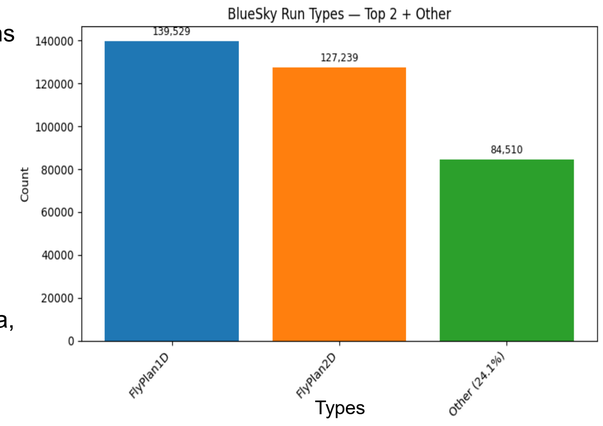
Results

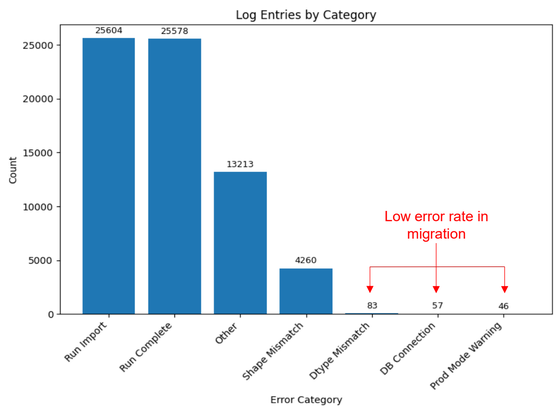


•**Growth Over Time:** BlueSky usage climbed from about 1-4k runs/month in 2021-22 to a peak of ~11k in October 2024 and has stayed strong at 6-9k/month into mid-2025.
•**Pandemic Impact:** COVID-19 shutdowns in 2020-21 drove run counts to near zero, with a rebound starting in late 2021 as in-person experiments resumed.
•**Daily Patterns:** By mid-2023, daily runs regularly reached 400-800 and early 2024 saw "super-run" days over 1,200, punctuated by routine zero-run gaps.
-Key Takeaway: BlueSky's dramatic growth and sustained high throughput underscore the need for proactive resource planning and meticulous metadata management.

This HXN-only chart shows that nearly **76%** of all runs on the Hard X-ray Nanoprobe use continuous "fly" scans:
•**FlyPlan1D-** the most common one-dimensional fly scan.
•**FlyPlan2D-**the go-to two-dimensional fly scan.
•**Other (24.1%)** – all remaining plans (diffraction, tomography, XANES, etc.) combined.

•**Takeaway:** On HXN, fly scans dominate (~76% of runs), so focusing on FlyPlan1D/2D covers most data, with "Other" reflecting a smaller set of specialized experiments.





This chart shows the breakdown of all parsed migration_log entries by high-level category:
Run Import (25,604) & Run Complete (25,578): Nearly every import is paired with a completion, dominating the log.
•**Other (13,213):** Miscellaneous messages-worth a quick audit.
•**Shape Mismatch (4,260):** The main warning, indicating frequent array-size fixes.
•**Dtype Mismatch (83), DB Connection (57),**
•**Prod Mode Warning (46):** Rare issues, each under 100 occurrences.
Key takeaway: Most entries record normal run imports and completions between MongoDB and PostgreSQL, while shape-mismatch warnings are the largest class of migration hiccups and should be the focus of any data-consistency improvements.

Conclusion And Discussion

Conclusion
Since its initial rollout, BlueSky usage at NSLS-II has exhibited a clear trajectory of recovery and growth. After a near-standstill during the COVID-19 shutdowns in 2021, runs per month steadily climbed from 1- 4k into the 4-6k range by 2023, peaking above 10k in April 2024 and reaching an all-time high of 1k in October 2024. Through mid-2025, monthly activity has remained elevated (6-9k runs), indicating that the facility now reliably supports substantially greater throughput than in its early years. Our analysis of run types further reveals that continuous "fly" scans (FlyPlan1D and FlyPlan2D) dominate roughly 80% of all experiments, with the remaining 20% split among mid-frequency and specialized routines. At the same time, our log analysis confirms that most entries-25,604 "Run Import" and 25,578 "Run Complete" messages-reflect successful migrations, while the next largest category, 4,260 "Shape Mismatch" warnings, highlights the primary area for data-consistency tuning.

Discussion
These patterns reflect both operational and scientific drivers. The 2021-2022 decrease corresponds to pandemic restrictions, while the rapid ramp in 2023-24 aligns with full adoption across beamlines and improved migration tools. Recurring spring and fall surges coincide with major experimental campaigns, and the daily burst-day behavior underscores peak-load challenges for data storage and analysis. The overwhelming prevalence of fly-scan plans suggests prioritizing infrastructure and metadata support for high-velocity data streams, while the long tail of specialized plans highlights opportunities to optimize lesser-used workflows or provide targeted user training.

Acknowledgement

I would like to thank my mentor, Yevgen (Eugene) Matviychuk, for their guidance and support throughout this project.