Navigating Exascale Operational Data Analytics: From Inundation to Insight

Woong Shin, Tim Osborne, Ahmad Maroof Karimi, Rachel Palumbo, Alex May,
Corwin Lester, Jesse Hines, Naw Safrin Sattar, Leah Huk, Scott Simmerman, Wesley Brewer,
Jeffrey Miller, Ryan Adamson, Olga Kuchar, Ryan Prout, Feiyi Wang, Scott Atchley, and Sarp Oral

Oak Ridge National Laboratory, Oak Ridge, TN

Email: {shinw, osbornetd, karimiahmad, palumborl, mayab, lestercp, hinesjr, sattarn, hukln, simmermansg, brewerwh, millerjl, adamsonrm, kucharoa, proutrc, fwang2, scott, oralhs}@ornl.gov

Abstract—In this paper, we address the challenges in achieving sustainable data-driven efficiency by providing a detailed exploration of the end-to-end operational data analytics (ODA) framework that evolved through two generations of supercomputer systems at the Oak Ridge Leadership Computing Facility (OLCF). This framework addresses large data streams ingested from heavily instrumented HPC environment that accumulates multi-terabytes per day. We outline the multifaceted data life cycle across HPC procurement, operations, and research & development, identifying key obstacles and design decisions that shape effective strategies in building and supporting data pipelines end-to-end. By sharing key insights and lessons learned from our experience, we offer recommendations for the HPC community on enabling sustainable operational data analytics and beyond. Our contributions aim to bridge the gap between potential and real benefits of operational data, guiding future efforts towards integrated and sustainable operational intelligence in high-performance computing environments.

Index Terms—Operational Data Analytics, HPC Post-Exascale Challenges, Monitoring, Telemetry, Data Analytics, Machine Learning Applications, Visual Analytics, Data Governance

I. INTRODUCTION

The increasing complexity of Exascale high-performance computing (HPC) systems, coupled with the diminishing performance gains per watt and the need for sustainability through carbon emission accountability, introduces significant operational challenges. Operational data is a crucial foundation in enabling continuous improvements towards reliable, safe, and efficient resource usage in the face of such challenges. Consequently, leading HPC sites are turning to operational data analytics (ODA) to navigate these challenges effectively [1]. The advancements in big data, data science, and machine learning across various domains offer promising opportunities for enhancing data-driven operational efficiency in HPC environments.

Despite the growing reliance on ODA to navigate the increasing complexities, significant barriers hinder its effective use. The volume of data generated and collected by HPC sites, presents a daunting challenge in managing and extracting actionable insights essential for enhancing operational efficiency and planning future systems. Additionally, disjointed efforts and a lack of cohesive long-term integration strategies further widen the gap between the potential benefits and the actual realization of the value of ODA.

There is an urgent need to address these challenges and shift towards a more integrated and effective ODA strategy in HPC. Data usage in HPC operations is lagging behind in contrast to other sectors [2] that thrive on data-driven methodologies such as internet search [3], cloud data centers [4], social-network services [5], ride-share industry [6], [7], and self driving [8], [9]. Nonetheless, there are opportunities to develop new strategies that account for the unique operational complexities of HPC.

In this paper, we describe the design and execution of an end-to-end operational data analytics (ODA) framework at an open science HPC facility, the Oak Ridge Leadership Computing Facility (OLCF). This framework serves as a centralized system for processing operational data from multiple supercomputer generations, handling 4.2 to 4.5 Terabytes of data daily across the HPC data center. It meets various organizational needs for operational data, from system administration to research and development, enabling holistic, data-driven strategies in addressing Exascale-era complications in HPC operations.

Throughout this exploration, we share our experiences and lessons learned from deploying this framework in an open science HPC user facility over two supercomputer generations, Summit and Frontier. This study provides insights and recommendations for navigating the unique challenges of implementing ODA within HPC environments. Our key contributions are as the following:

End-to-end data life cycle in the context of HPC operational efficiency: We report the multifaceted, end-to-end nature of the data life cycle within an HPC organization that supports generations of HPC systems. Teams from various domains contribute different viewpoints and use cases, leading to multiple phases for the same data source.

End-to-end framework for ODA: We detail the implementation of the framework to facilitate sustainable ODA

TABLE I
AREAS OF OPERATIONAL DATA USAGE IN A HPC ORGANIZATION

System Management	
System Administration	System performance, stability and reliability ensurance: compute, interconnect, storage
Facility Management	Reliable and energy efficient power and cooling supply system design and operations
Cyber Security	Detection, diagnosis and prevention of security issues
Operations	
User Assistance	Diagnostics for swift troubleshooting and solutions
Administrative	
Program Management	Resource allocation, coordination, and reporting to sponsors
Job Scheduling	Job execution priority adjustment based on program needs and user requests
Procurement	
System Design	Technology integration, tuning, testing, and projection for future systems
R&D / Cross Cutting Thrust Areas	
Performance	Performance optimization, tuning
Reliability	Reliability projection and prediction
Applications	Runtime performance monitoring and optimization, tuning, energy efficiency
Energy Efficiency	Energy usage optimization from various layers of an HPC data center

from data ingestion to application. Our discussion covers the technical frameworks employed, alongside the essential policy and life cycle enhancements that streamlined data utilization and adoption across the organization.

Recommendations to the community towards the future: Through our exploration of data life cycle and support infrastructure, we share key insights from the process of enabling ODA across two high-ranking, large-scale HPC systems.

II. BACKGROUND AND MOTIVATION

A. Operational Data Analytics from Large-Scale HPC Sites

Amidst the pursuit of exascale computing, there is a notable shift towards ODA marking a transition from traditional monitoring to a more integrated approach in handling operational data [1]. ODA represents an evolved framework that supports a common infrastructure, enabling organization-wide data collection, engineering, and distribution [10]. This paradigm not only caters to the three Vs of Big Data—volume, velocity, and variety—but also facilitates a wide array of use cases through its capabilities [1], [11]–[16]. Such systems have sparked significant interest in new, data-driven operational improvements and have opened avenues for research innovations, particularly the application of machine learning algorithms on operational data [17]–[21].

B. Operational Data at OLCF

The Oak Ridge Leadership Computing Facility (OLCF) serves the DOE Office of Science by providing high-end HPC systems for large-scale computational tasks that address major research areas like advanced scientific computing, basic energy, biological and environmental, fusion energy, high energy physics, and nuclear physics. Managed and operated by a dedicated HPC organization, the OLCF plays a pivotal role in propelling scientific progress with its generations of

top-tier HPC systems. Each system debuted at the topmost ranks of the Top500 list [22].

Over its two-decade journey towards exascale computing, the OLCF has faced challenges in heterogeneity, scale, and operational complexity. Operational data is used to meet these challenges and support its mission to advance scientific knowledge. Table I illustrates the multifaceted usage of operational data by the HPC organization which serves many purposes: addressing operational demands and supporting HPC research objectives aligned with its mission. The diverse use of data is powered by data streams emitted from many parts of the organization, including data from scientific workloads and usage patterns that drive generations of HPC systems.

C. Challenges in Making Operational Data Work for Us

Heterogeneous, continuous use of streamed data in the organization requires new perspectives on operational data. Data is immediately valuable upon creation, necessitating that data consumers process it in real-time, rather than waiting for the entire dataset upon system decommission. Moreover, the data outlives its originating system and is crucial for planning future generations of supercomputers. The increasing size, complexity, and diversity of these data streams render traditional management approaches by system administrators alone as a side-job impractical.

The proliferation of open-source tools has significantly improved data collection [1]; however, the operationalization of this data across organizations remains a substantial challenge. Despite abundant data acquisition, there is a notable gap in the end-to-end understanding of how the data is used, resulting in the accumulation of unused data and uncoordinated efforts that fail to deliver the right kind of data for operational use. Technical issues, such as inadequate tools for diverse data usage needs and difficulties in making large, complex datasets accessible, exacerbate these challenges. Additionally, non-technical barriers like policy enforcement and compliance further delay the process from data ingestion to actionable insights.

III. OVERVIEW

In addressing such challenges, our overarching aim is to elevate the use of operational data within the supercomputing life cycle, treating it as essential rather than optional in an HPC organization. We focus on the end-to-end considerations of ODA to ensure that every facet of data utilization is optimized, empowering the entire organization in the process. The following are the major considerations we made in this end-to-end effort:

- **Reliable source of data:** singular, trusted, stable location for managing operational data in the organization.
- **Streamlined data acquisition and usage:** systematic engineering, policy, and life cycle support for data acquisition and usage.
- **Accelerated, heterogeneous use of streaming data:** accelerating data life cycle by consolidation of high-impact
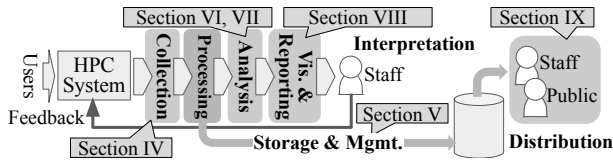
Fig. 1. End-to-end of the data life cycle is formed around operational control feedback loop handling data streaming from the HPC system

Fig. 2. Sensor data stream establishment and its usage is driven by use cases and matures over time (i.e., L0 to L5) impacting future operations

Sources / Areas	System Mgmt.	User Assist.	Facility Mgmt.	Cyber Sec.	Apps.	Prgrm Mgmt.	Procure-ment	R&D
Perf. counters					L0 L0		L0 L0	L0 L0
Resource util		L0 L0			L0 L1	L5 L5	L2 L1	L0 L1
Power & Temp.	L1 L1	L0 L3	L4 L4		L2 L2		L1 L1	L5 L3
Storage client	L1 L1	L5 L5			L0 L1		L2 L1	L5 L1
Interconnect client	L1 L1	L5 L5			L0 L1		L2 L0	L0 L1
Storage System	L4 L2						L2 L0	L0 L0
Interconnect	L0 L0	L0 L0					L2 L1	L0 L0
Syslog & Events	L5 L5	L5 L5	L4 L1	L5 L4			L4 L2	L4 L1
Resource Manager	L5 L5	L5 L5		L5 L4	L5 L5	L5 L4		L5 L3
CRM		L5 L5			L5 L5	L1 L1		
Facility			L5 L4				L5 L5	L4 L3

(left: Compute System)

Fig. 3. Data requirement, development in various areas: maturity of data usage across systems in each cell (left: Mountain, right:Compass) expressed in L0 ∼ L5 as in Figure 2. Boldface outline is where the teams behind an area (X-axis) is responsible of producing the type of data (Y-axis) as the owners.

data identification, data management, data processing, and data publishing efforts.

- **Sustainable, reliable, and reproducible operational impact:** enable organizational data usage through well-canned applications and services implemented on top of repeatable, sustainable industry best practices.

These considerations focus on refining a unique, yet widely applicable, data life cycle in HPC ODA (Figure 1). This life cycle centers around a manual operational feedback control loop. This loop is powered by batches of data generated from real-time data streams, which are then shared internally within the organization and externally with the HPC community. Sections IV to IX explore the challenges and insights of each key stage, aiming to enhance our understanding of HPC ODA and identify opportunities to accelerate the iterations.

IV. DATA COLLECTION

Data collection efforts are driven by topical Subject Matter Experts (SMEs) or project Principal Investigators (PIs) collaborating with system owners with the goal of enhancing performance, reliability, and energy efficiency. The process for developing data streams is depicted in Figure 2. This process begins with a data collection plan informed by experiences and use cases with prior systems very early in the process. This plan involves engineering and refining the collection process for broader operational use.

A. Multi-source Multi-use Nature of Operational Data

Data collection faces challenges due to the prototypical nature of our systems. Securing vendor cooperation for un-planned sensor implementation can be difficult. Additionally, there is a trade-off between minimizing system overhead and ensuring the quality of signals and features at scale. This balancing act is often constrained by the available technologies for data extraction and delivery. Vendor-supplied technology plays a crucial role often requiring iterative communication between subject matter experts, system administrators, and vendors to enable data streams that are sufficiently usable downstream but within acceptable overheads of the system. These issues are mitigated by leveraging lessons from previous generations and optimizing sensor data collection for future systems, but consequently, takes time due to end-to-end trial and error.

Figure 3 illustrates the current status of data stream development throughout the organization for the past two generations of supercomputers expressed in terms of the stages depicted in Figure 2 as a degree of data usage readiness at the current
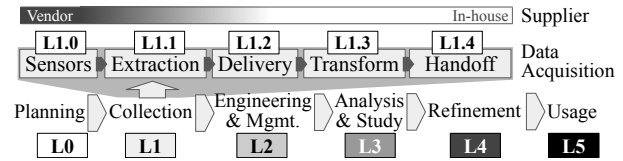
date. Critical data sources (X-axis) are primarily generated in the system management area to meet immediate operational requirements, yet these data streams hold value for numerous other areas as well (Y-axis). Despite the necessity for multiple data streams outside of system management, there remains a gap in achieving the full readiness and utility of these datasets across various organizational domains.

B. SME or PI Driven Data Stream Development

In addressing these challenges, our approach is to have topical SMEs across different areas proactively ensure that necessary capabilities and technologies are considered early in procurement processes. These activities include the robust development and demonstration of use cases outside the immediate system management scope, directed both at system owners and the vendor community. It also includes identifying and embracing innovative approaches in data extraction and delivery mechanisms (i.e., monitoring infrastructure) to ensure the delivery of sensor data is guaranteed outside of the system. This is especially true for data collection which can be too invasive to the system. New approaches such as fully leveraging the out-of-band data sources [23] via the management network [24], [25] or leveraging per-job instrumentation based on technologies such as Darshan [26] has been successfully employed to mitigate such issues.

Lessons Learned: Data stream development is a non-trivial effort that should be planned ahead and coordinated by capturing the future and current use cases and needs from various parties in the organization. To guarantee success, responsible topical SMEs or PIs should engage

Major Data Streams	Data Ingest
Compute storage client	~2GB/d
Compute Power & Temp.	~537GB/d
Storage System	~3.3TB/d
Interconnect	~32GB/d
Syslog & Events	~8.64GB/d
Resource Manager	~11MB/d
CRM	~350MB/d
Facility	~2.5MB/d

a) Ingest rate of major data streams
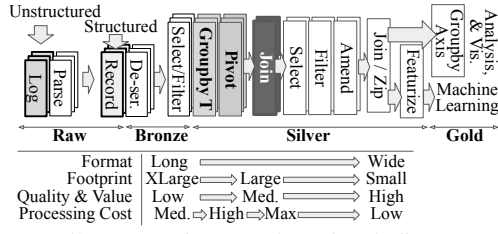
b) Anatomy of ODA queries or data pipelines
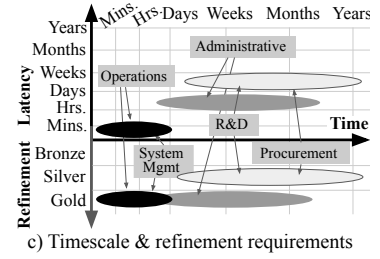
c) Timescale & refinement requirements

Fig. 4. Raw data ingest rate from our current system can go up to terabytes scale per day - a) Data pipelines handling large scale multiple data streams show a common pattern - b) Implementation of the pipelines are driven by the multi-timescale data usage - c)

with vendors as early as the prior generation of the system and ensure that data stream capabilities exist in the vendor technology offering and see through the delivery of data.

V. DATA ENGINEERING AND DATA MANAGEMENT

The matrix between the multifaceted use of data streams from various areas (Figure 3) illustrates the complex relationship between data producers and consumers. To manage this complexity, we have adopted an hourglass-type architecture that introduces a multi-tenanted, centralized data management and engineering infrastructure. Based on a deep understanding of the end-to-end process of data stream development to data artifact usage, this infrastructure has evolved into an optimized, flexible, self-service, one-stop shop for various staff projects that need data, storage, and the compute power to handle the three Vs of big data.

A. Large Data Flows and the Anatomy of ODA Data Pipelines

Figure 4-b) illustrates the common anatomy of ODA queries or data pipelines conceptually broken down in terms of SQL clauses regardless of the actual implementation. Through the pipeline stages, data is refined through "Bronze", "Silver", and "Gold" states as an adaptation of the "Medallion Architecture" [27]. Initially, raw data undergoes a transformation into a tabular long-format, where each row encapsulates an individual sensor observation, marking the preliminary "Bronze" stage of data refinement. Subsequently, this dataset is aggregated over designated time intervals (e.g., every 15 seconds) to reconcile differences in sample rates and then pivoted into a wide format. In this format, each row signifies a specific component or node, potentially integrated with additional datasets (such as job allocation logs) for contextualization and further refinement. This results in the "Silver" stage of refinement, characterized by its more processed nature. The ultimate phase includes: slicing and dicing the data through group-by aggregations relevant to specific analyses; visualizations intended for human interpretation; data prepared for machine learning model training through featurization—yielding "Gold" stage data artifacts that represent the final refined product.

Distribution of these stages is heavily influenced by the *control loop* timescale of an operational domain and its corresponding data refinement requirements, which dictate the pipeline latency constraints (as depicted in Figure 4-c)). This dynamic is especially important in scenarios involving

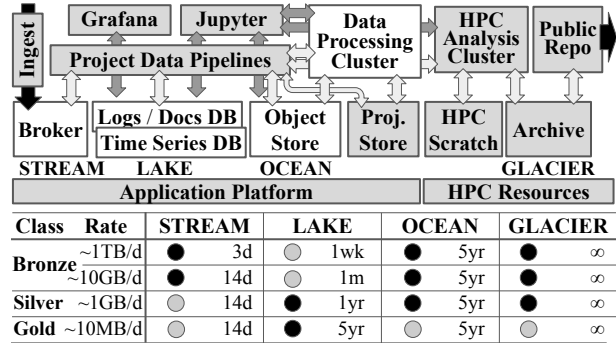| Class	Rate	STREAM		LAKE		OCEAN		GLACIER	
Bronze	~1TB/d	●	3d	◌	1wk	●	5yr	●	∞
	~10GB/d	●	14d	◌	1m	●	5yr	●	∞
Silver	~1GB/d	◌	14d	●	1yr	●	5yr	●	∞
Gold	~10MB/d	◌	14d	●	5yr	◌	5yr	◌	∞

Fig. 5. Common data services (white) and project specific resources (gray) managed in a tiered fashion. Each tier (STREAM, LAKE, OCEAN, GLACIER) focuses on different classes [27] of data artifacts (left: darker circles) with class specific retention time (right: duration). Project resources are "outsourced" (gray boxes) from HPC and support resources from the organization (i.e., App. platform & HPC resources).

large data flows where "Bronze" stage artifacts seldom serve immediate analytical purposes without undergoing substantial transformations—entailing a series of group-by aggregations, pivots, and joins that necessitate considerable I/O operations and data movement to achieve a more compact and computationally efficient "Silver" stage.

B. Data Services, Technologies, and Infrastructure

Figure 5 outlines the architecture of data services and resources that streamline the per-project journey from data ingestion to archival or public release. The architecture facilitates broader utilization of refined datasets, either through continuous streams in the STREAM service (data streaming broker) or via ever-appended parquet-based highly compressed tabular data in the OCEAN service (S3 object store). For long-term preservation, data is archived in the GLACIER (Tape Archive), whereas immediate real-time usage needs are catered to by the LAKE (online database access) service. Additionally, downstream artifacts are disseminated through a site-wide public data repository Constellation [28], [29], ORNL's public data repository, following a formal approval process outlined in Section IX.

Each building block and underlying technology was chosen for efficient data access, availability, and resource usage. Apache Kafka [30] serves as a core component, acting as FIFO buffers for in-flight data in distributed multi-project pipelines. ElasticSearch [31] and Apache Druid [32] are used for real-time diagnostics and debugging, targeting unstructured and

time series data, respectively. For long-term storage demands, Apache Parquet [33] with MinIO [34] offers a column-oriented compressed file format, ensuring significant data compression and minimal I/O footprint. Apache Spark [35] structured streaming [36] is adopted for high-volume processing of multiple data streams, providing SQL-based real-time processing along with advanced failure and recovery mechanisms that can be difficult to re-engineer from scratch.

C. Project Specific Data Engineering

Supplementing these centralized services, application platform resources play a crucial role in providing access to compute power, memory, and temporary storage necessary for running sophisticated data pipelines. Our platform, called Slate [37], is constructed atop Kubernetes [38] (OpenShift [39]) offering a non-HPC resource environment designed for applications requiring continuous uptime like databases, web server data portals, message queue software, or stream processors. This self-service environment empowers project subject matter experts to construct and manage their data pipelines autonomously, leveraging project-specific allocations to meet their unique requirements while maintaining our multi-tenant security model for the workloads running on the platform.

For more demanding computational tasks that exceed the capabilities of Slate's non-HPC resources, projects have the option to tap into high-performance computing (HPC) systems. These powerful platforms support large-scale batch processing tasks such as data amendment operations, backfills, or extensive analysis campaigns by utilizing allocated modest amounts of node hours from project allocations. This approach, similar to a Platform as a Service in the cloud computing space, enabled us to coordinate the compute, memory, and storage usage of multiple projects in an efficient way enabling higher utilization of physical resources.

Lessons Learned: Data services and infrastructure can be optimized by identifying common patterns of the data pipelines and their artifacts. In this process, known data management techniques from the database community and the big data community play a significant role in bounding resource usage. Implementing tiered data management, data reuse, compression, SQL interfaces, failure and recovery mechanisms, and stream processing all made a huge difference. Project-specific allocations were effective in supporting multiple projects that required storage and compute resources. Common data services bound overall resource usage by eliminating redundant work.

VI. DATA DISCOVERY, EXPLORATION AND ANALYSIS

Once data is collected and made accessible to the organization, it requires rigorous exploration to be understood and utilized effectively. Data discovery and exploration is initiated by SMEs or staff project PIs aiming to understand the quality, meaning, and value of a pile of relevant raw datasets that sometimes have a footprint of terabytes. Challenges such as limited information during the data discovery phase and the difficulty of processing backlogs of undiscovered or unrefined data can significantly delay subsequent phases, leading to low data coverage.

A. Data Exploration Campaign

To navigate these complexities, we initiate "data exploration campaigns" focused on breaking new ground into a set of datasets related to an operational topic relevant to the mission. Aiming for developing a sustainable pipeline for large-scale data streams, these path-finding activities concentrate resources to address various challenges once and for all for the organization.

These data exploration campaigns first focus on building a data dictionary that has qualitative information about the dataset such as sample rate, failure rates, logical and physical sensor location, and their meaning with respect to the underlying process or system. Here, the role of the system provider and vendor is crucial, but this area has room for improvement. This often involves costly interactions with the vendor tracking down the engineers responsible for developing the sensors and acquiring authoritative knowledge due to the bleeding edge nature of the hardware.

B. Developing High-Impact Upstream Data Pipelines

Recognizing the complexity of managing terabyte-scale data streams, initial efforts focus on identifying and refining the processes necessary to transform raw data (Bronze state) into a more usable form (Silver state). To manage this efficiently, we "outsource" these tasks onto the HPC system or a dedicated data processing cluster adopting distributed data processing frameworks like Apache Spark. This burst of activity is driven by a goal of identifying the costly data transform phases and implementing upstream data stream processing units to precompute refined Silver datasets in real-time. This transition from batch to stream processing amortizes the cost of refining datasets over a long period of time while making refined datasets that significantly accelerate iterations of downstream activities.

A deeper understanding of the intricate processes involved in refining datasets across various timescales and tiers helped developing strategies to mitigate the pressure in data collection and management. For example, terabyte-scale Bronze datasets can be stored in cold storage in a frozen state (GLACIER) as there was very little value in serving unrefined data sets in hotter data tiers until upstream data pipelines are developed creating refined datasets that are manageable.

C. Drivers of Data Exploration Campaigns and their Impact

User support and program needs are essential drivers of data exploration campaigns, focusing on enhancing user experience and streamlining reporting processes to sponsors. These campaigns leverage a broad spectrum of datasets to create intuitive dashboards for easy access to information. Based on experiences supporting users on previous systems, such projects initiate explorations to pinpoint which data streams

and metrics are most relevant. Further aims to identify analysis methods to best serve the diagnostic processes.

R&D activities also play a significant role in such campaigns aiming to understand system responses under user behaviors, encompassing performance, reliability, and energy efficiency. Heavily relying on profiling user application activities, R&D initiatives process terabytes of data, condensing them into a manageable format for deeper analysis. This process lays the groundwork for developing use cases and refining data analysis techniques.

System design and procurement decisions also motivates explorations on operational data to determine the specifications of new supercomputing systems. This involves a meticulous balance between various system components such as compute capacity, memory, bandwidth, and storage within the confines of budgetary limitations. A data-driven approach, grounded in the analysis of long-term telemetry datasets reflecting user behavior, ensures that procurement decisions are made with precision.

> **Lessons Learned:** The primary bottleneck in HPC operational intelligence lies within the initial stage of large-scale stream exploration—directly impacting overall data coverage and usage. Consolidated data exploration campaign efforts toward sustainable pipeline development play a crucial role. Strategic investments in enabling streams of high-coverage sustainable data artifacts are pivotal to unlocking downstream high-profile innovations and operational impact.

VII. Data Visualization and Reporting

Data visualization and reporting are crucial for empowering day-to-day operations by transforming raw data into actionable insights which demands a broad combination of technical expertise as it forms the culmination of the data pipeline at the end, where the real empowerment happens. Despite its importance, this area is often undervalued, seen as a task for interns rather than a critical operational function. Key challenges include reducing delays from data ingestion to visualization while ensuring low-latency interactivity for end-users and achieving higher data coverage without overwhelming the analysis process. Time to delivery and ensuring end-to-end sustainability also pose significant obstacles, requiring mature data pipelines developed through concrete steps.

A. Sustainable "Well Packaged Data Applications"

In addressing these challenges, our approach focuses on creating purpose-driven applications crafted through well-studied analysis, models, and interactions. Coordination with prior data life cycle stages mentioned in prior Sections are crucial. Empowerment of field operators, system administrators, and user support teams should be one of the focus areas in developing data streams, analysis methods, and their data pipelines.

Due to the large lead-time in enabling such teams and efforts, we found it best to design and optimize towards serving generations of systems through disciplined continuous improvement of such services potentially all the way from research to production. Our approach in well packaged data applications as software services resulted in several long-standing application that takes advantage of the real-time holistic data-driven view to empower major operational activities and thrust areas.
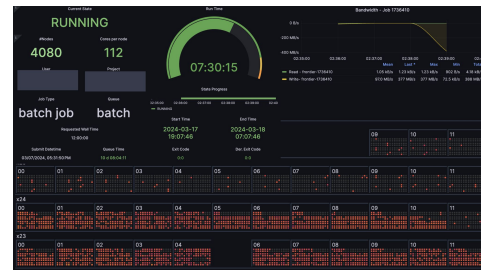
B. Data Powered Software Services

Fig. 6. User assistance: increases productivity of issue diagnosis by providing easy access to various system metrics and job oriented metrics

User assistance dashboards: The User Assistance (UA) group uses specially designed dashboards to improve handling of daily user tickets (Figure 6). These dashboards compile data from various sources, including compute, storage, and system logs, all integrated with job node allocation details for a comprehensive overview. This type of compilation replaces the old method of manually checking different systems or consulting with experts, streamlining the problem-solving process through clear visualizations tailored for quick issue identification. As a result, the group has seen a significant decrease in the time it takes to resolve user problems, making the troubleshooting process more efficient and effective for everyone involved.
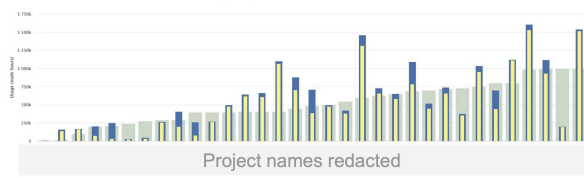

Fig. 7. Screenshot from RATS-Report showing project usage (CPU vs. GPU) across an allocation program which is easily accessed in real-time

User Resource Usage Report: RATS Report: Figure 6 depicts RATS-Report, the central reporting infrastructure for the HPC user facility, offering comprehensive insights into usage data such as node-hours on compute resources and filesystem storage utilization. This system provides users access to over a decade's worth of utilization data, supporting customized visualizations for diverse metrics including resource usage, project allocations, and user activity. A key feature is its capability to track burn rates for project allocations, aiding in efficient job scheduling. Daily data ingestion encompasses a vast range of sources including compute job logs from multiple

schedulers, GPU stats, and filesystem usage logs, amounting to potentially millions of parsed log lines.
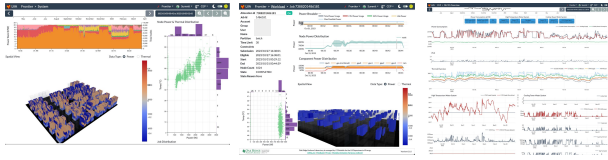

Fig. 8. Live Visual Analytics (LVA) provides near real-time low latency interactivity into years worth of high-dimensional power and thermal profile data (left: system view, mid.: job allocation, right: cooling plant)

Power and Energy Analytics: To address the challenge of managing and analyzing extensive high-dimensional datasets in HPC energy efficiency, a custom interactive visual analytics service, Live Visual Analytics (LVA), was developed (Figure 8). LVA facilitates rapid exploration of years of accumulated power profiling data, despite the high volume of incoming data (e.g., 0.5 TB/day for the Frontier supercomputer). This capability is enabled by a specialized data refinement pipeline that delivers contextualized job power profiles, which vastly reduces the amount of processing required in interactive queries from the user interface.

Cybersecurity: Copacetic: High-resolution system information such as performance counters and system telemetry Extreme-scale data analytics capabilities are essential to the integrity of scientific computing. [40] [41] *Copacetic* is an in-house developed analytics tool that requires a reliable feed of real-time events and logs from non-homogeneous data sources provided by ODA infrastructure. It detects when certain specific combinations of network availability, system state, and user behavior occur and informs administrative teams to take security-specific actions. ODA infrastructure greatly lowers the barrier of access to data sources in a way that is otherwise impossible for traditional batch-based security information event monitoring (SIEM) tools.

> **Lessons Learned:** Ensuring that the value of data effectively reaches day-to-day operations through usable visualization and reports demands special care and consideration. Adopting a "sustainable software service" model worked well in achieving quick time-to-delivery while addressing specific operational needs through customizable applications integrated into workflows.

VIII. MACHINE LEARNING (ADVANCED DATA USAGE)

Machine learning (ML) enhances operational data analytics by automating the approximation of system dynamics through models derived from large datasets. These models serve as tools for dimensionality reduction aiding in descriptive or diagnostic analytics, and act as proxies for the actual system, enabling predictive or prescriptive analytics through forecasting and optimization. Despite its potential and demonstrated benefits in research [17]–[21], the application of ML in operational data analytics remains challenging. We identified that this is largely due to difficulties in formulating problems
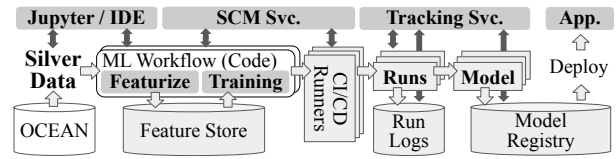

Fig. 9. Per-project implementation of a machine learning pipeline for repeatability and reproducibility

and integrating machine learning outcomes and workflows into existing engineering processes.

A. Problem Formulation

Our approach in addressing problem formulation is to drive the process with an overarching understanding of the requirements, constraints, and challenges of HPC operational data analytics. We have found that many of our use cases could be understood by modeling the continuous improvement loops as manual feedback control loops at a relevant timescale. This control loop operates on time series data or events that continuously stream in large volumes and velocity which humans need to understand deviations, make optimal decisions based on the understanding of the system, and make adjustments within the required time scale. Such an effort can be challenging due to the interdisciplinary nature of understanding both the HPC operational domain and ML. Even with experts in each area, basic training on HPC or ML which each individual lacks was necessary for basic communication in a team setting.

With this conceptual view, we can identify robust impact points and constraints that we can utilize to define inputs, outputs, constraints, and goals of a machine learning problem. Further, identify major machine learning challenges unique to operational data analytics which we focus our efforts in solving. Progress in ML use cases is currently heavily bottlenecked by the data itself due to its streamed, skewed, and lossy nature which starves ML development iterations with unknown future data, low-yield features, rare events, and missing data. This often results in upstream data stream and pipeline refinements that are costly or sometimes impossible. Data quality enhancements and the advancement of ML techniques that can cope with these issues are equally important. These factors should be also accounted into problem formulation.

B. Machine Learning Engineering

To facilitate repeatable, reproducible ML model development and usage targeted in our operations, we focus on developing ML pipelines by carrying out such projects similar to a software engineering project by focusing on code management with some extensions to handle data and models, as illustrated in Figure 9. This process starts from importing Silver class refined batches of datasets on OCEAN, managing featurized data through version-controlled project feature stores (DVC [42]), employing CI/CD workflow support in private deployment of software change management services (GitLab [43]) for training orchestration, and tracking experiments and distributing models via an ML tracking service (MLflow [44]) for downstream inference workloads. To maintain sufficient flexibility

in this process, such pipelines are managed and developed through project-specific allocations. Due to the reduced size and refined nature of the Silver class datasets, most of our training workloads are expected to fit in non-HPC resource allocations.

C. ODA Applications of Advanced Data Usage

Machine learning and advanced data usage in our facility are driven by the conceptual operational feedback control loop we have identified. On top of the aforementioned ML engineering project environment, these use cases are currently work in progress aiming for tangible operational impact. Many of our use cases found place in the context of HPC energy efficiency due to its end-to-end nature.
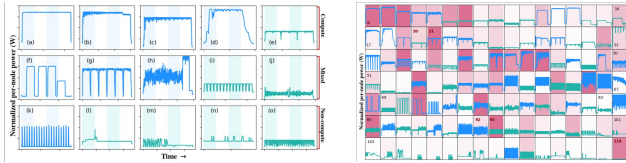

Fig. 10. Profiling jobs based on their power profile (left). A neural network-based classifier automatically groups power profiles based on their similarities (right) — cells are profile shapes and the color is the observed population.

User Job Profiling on Power and Energy: In the context of energy efficiency, a novel real-time job classification pipeline [45] enhances analysis by clustering job power profiles based on their similarity in consumption patterns using a neural network (Figure 10). This classification not only facilitates easier navigation through the data but also offers insights into the relationship between application resource use and its overall energy impact on the system.
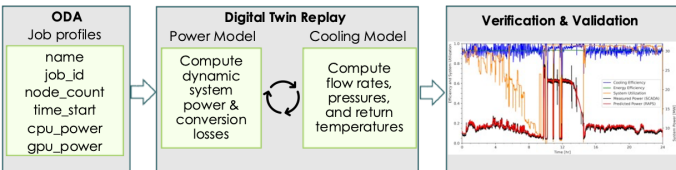

Fig. 11. Telemetry flow and input of ExaDigiT (left), the telemetry replay of a HPL run on the simulators (middle) and the virtual cooling system response (right) during verification and validation.

Digital Twins: ExaDigiT: We have developed a comprehensive digital twin of the Frontier supercomputer called ExaDigiT [46]. The digital twin has three main modules: (1) a resource allocator and power simulator, (2) a transient thermo-fluidic cooling model, and (3) a virtual reality model of both the supercomputer and central energy plant. Such a twin can be used to study "what-if" scenarios, system optimizations, and virtual prototyping of future systems. The system replays various telemetry data from the HPC data center for verification and validation of the power and thermo-fluidic models. As white-box models based on thermodynamics, these models overcome the limitations of black-box data-driven machine learning models that do not extrapolate to unknown

TABLE II
CONSIDERATIONS FROM THE ADVISORY CHAIN

| Consideration	Description
Data Owner	Considers purpose and potential interpretation of the data that can harm ongoing operations.
Cyber Security	Prevent leakage of PII data embedded within the data or information that can identify certain projects or users.
Legal	Provides guidance on legal requirements defined by contractual obligations as well as any national regulatory concerns.
Institutional Review Board (IRB)	Federally mandated entity that oversees the protection of human subjects in research ensuring rights and welfare of human research subjects are protected [47].
Management	Organizational approval on publications or artifacts reviewing alignment with the facility mission.
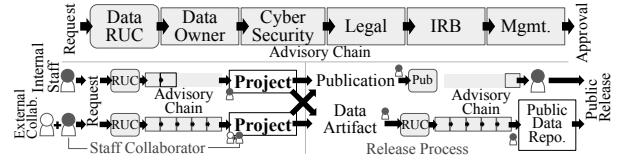
Fig. 12. Data distribution workflow for internal distribution and collaborations with external personnel. Publications and data artifacts are released with approvals from the advisory chain.

states. The framework reveals the complex transient dynamics of the cooling system, handles synthetic or real workloads, and predicts energy losses due to rectification and voltage conversion (Figure 11-right).

Lessons Learned: Beyond the hype, the status quo of ML in HPC ODA is well beyond being a *hammer in search of a nail* but it requires interdisciplinary effort to integrate it into existing engineering efforts and make meaningful impact. Even with established experts in each domain, HPC operations and ML, basic cross-domain training was required for the necessary communications between experts to pursue such efforts. Repeatable and reproducible execution of machine learning development is crucial. We have found practices in software engineering useful as a starting point as those handle code and can be reasonably easily extended to data and models.

IX. DATA GOVERNANCE AND MANAGEMENT

In producer-consumer relationship matrices across many sources and usage areas, distributing data can be challenging without proper data governance and management support. The power and value of operational data is important to a variety of stakeholders; that is, system owners and data owners at the organization level, the funding agency, government, vendors, and the hosting institute which is subject to a variety of rules, regulations, contracts, and laws. While it is extremely difficult for individual projects to navigate such a high-stakes environment, the process serves as a standard mechanism that can facilitate the process safely and minimize delays.

A. Data Access, Governance and Management

Moving successfully through any facility-wide workflow requires coordination and collaboration between multiple entities. Figure 12 shows steps we take to distribute the data while ensuring safety considerations. This is done by reviewing every data usage request through an advisory chain described in (Table II). The review starts either from internal staff either for their projects or on behalf of an external collaboration (e.g., university collaboration) submitting a request to a data resource usage committee (DataRUC) each at the stage of starting a project or optionally at the phase releasing publications or data artifacts to a wider audience via appropriate channels. In particular, for datasets, the data is curated, and archived in a public repository for public usage.

B. Data Usage

With the approvals from the advisory chain, access to the data is provided and tracked via various channels suitable for the projects in a fine-grained manner. Internal projects are provided access to data service resources (Figure 5) such as STREAM, LAKE, or OCEAN to acquire access to relevant data to either (1) enable data visualization and reporting applications (STREAM, LAKE) or (2) carry out a historical analysis campaign to publish a paper or further develop pipelines (OCEAN). In cases of external collaborations, data is released in an approved project area or allocation provided by internal staff in the form of files or project database accesses. In the process of acquiring approvals for external data usage, internal staff hosting such projects carry out data sanitization or anonymization tasks with the guidance of the curation and cybersecurity staff before the data reaches external users.

Towards the HPC community, this data distribution process accelerated the safe release of Summit's power and energy data [48], GPU failure data [49], I/O data (Darshan) [50], [51]. This process was also applied to the release of Frontier's 2023 June HPL run submission data [52].

> **Lessons Learned:** Having a comprehensive approval process and gateway may sound counterintuitive toward the goal of organizational empowerment. However, we have found such a process is instrumental in accelerating empowerment.

X. CONCLUSION

In this paper, we have explored the end-to-end framework of ODA within an HPC organization, pinpointing crucial investment areas that helped us advance ODA to support a multi-tenanted and multifaceted use of operational data. Our journey with ODA through the recent two generations of large-scale supercomputers has uncovered significant insights. These not only shed light on the changing dynamics of data use in HPC operations, which makes ODA challenging but also pave the way for identifying cost-effective solutions for handling complexity and strategies for maintaining operational impact with high data coverage.

The transition from traditional monitoring to ODA is driven by the complexities of big data and versatile, multi-purpose data streams that feed real-time operational feedback loops in a matrixed producer-consumer environment across various time scales. This shift makes ODA resource-intensive; however, a thorough understanding of its full data life cycle allows for cost-effective infrastructure development and process optimization. Additionally, sustainable pipelines and software services, along with policies and workflows for data distribution, are required to fully empower the organization. The role of SMEs (Subject Matter Experts) and PIs (Principal Investigators) is crucial from end to end. These approaches not only lower operational costs but also enhance data coverage and impact.

In the context of HPC ODA, there is a challenge of achieving immediate data availability in the face of the relatively short lifespan of supercomputers. To address this challenge, we stress the importance of minimizing re-work by consistently investing in infrastructure and accumulating knowledge across different system generations. We also advocate for standardized HPC use cases that could speed up vendor-driven enhancements in sensor documentation and data collection technologies.

Looking forward, as we move beyond exascale computing into an era marked by sustainability challenges, complex scientific workflows, and multi-facility campaigns [53], the evolution of HPC operations will increasingly depend on supplementing human decision-making with sophisticated operational data analytics. Proficiency in data-driven operational intelligence will become even more essential in this new era.

REFERENCES

[1] M. Ott, W. Shin, N. Bourassa, T. Wilde, S. Ceballos, M. Romanus, and N. Bates, "Global Experiences with HPC Operational Data Measurement, Collection and Analysis," in *2020 IEEE International Conference on Cluster Computing (CLUSTER)*, Sep. 2020, pp. 499–508.

[2] Nathan Benaich, Alex Chalmers, Othmanne Sebbouh, and Corina Garau, "State of AI Report 2023," 2023. [Online]. Available: https://www.stateof.ai/2023-report-launch

[3] S. Levy, "How Google is Remaking Itself as a "Machine Learning First" Company | Backchannel," *Wired*, Jun. 2016, section: tags. [Online]. Available: https://www.wired.com/2016/06/how-google-is-remaking-itself-as-a-machine-learning-first-company/

[4] Rich Evans and Jim Gao, "DeepMind AI reduces energy used for cooling Google data centers by 40%," Jul. 2016. [Online]. Available: https://blog.google/outreach-initiatives/environment/deepmind-ai-reduces-energy-used-for/

[5] Aarti Basant, "Scaling data ingestion for machine learning training at Meta," Sep. 2022. [Online]. Available: https://engineering.fb.com/2022/09/19/ml-applications/data-ingestion-machine-learning-training-meta/

[6] Zoubin Ghahramani, "Uber AI in 2019: Advancing Mobility with Artificial Intelligence," Dec. 2019. [Online]. Available: https://www.uber.com/en-GB/blog/uber-ai-blog-2019/

[7] N. Kankani, "Scaling AI/ML Infrastructure at Uber," Mar. 2024. [Online]. Available: https://www.uber.com/blog/scaling-ai-ml-infrastructure-at-uber/

[8] Mark Harris, "Tesla's Autopilot Depends on a Deluge of Data," *IEEE Spectrum*, Aug. 2022. [Online]. Available: https://spectrum.ieee.org/tesla-autopilot-data-deluge

[9] Tesla, "Tesla AI Day 2021," Aug. 2021. [Online]. Available: https://www.youtube.com/watch?v=j0z4FweCy4M

[10] A. Netti, W. Shin, M. Ott, T. Wilde, and N. Bates, "A Conceptual Framework for HPC Operational Data Analytics," in *2021 IEEE International Conference on Cluster Computing (CLUSTER)*, Sep. 2021, pp. 596–603.

[11] A. Netti, M. Müller, C. Guillen, M. Ott, D. Tafani, G. Ozer, and M. Schulz, "DCDB Wintermute: Enabling Online and Holistic Operational Data Analytics on HPC Systems," in *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing (HPDC)*, Jun. 2020, pp. 101–112.

[12] A. Borghesi, A. Burrello, and A. Bartolini, "ExaMon-X: A Predictive Maintenance Framework for Automatic Monitoring in Industrial IoT Systems," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 2995–3005, Feb. 2023, conference Name: IEEE Internet of Things Journal.

[13] A. Netti, M. Ott, C. Guillen, D. Tafani, and M. Schulz, "Operational Data Analytics in practice: Experiences from design to deployment in production HPC environments," *Parallel Computing*, vol. 113, p. 102950, Oct. 2022.

[14] M. Terai, K. Yamamoto, S. Miura, and F. Shoji, "An Operational Data Collecting and Monitoring Platform for Fugaku: System Overviews and Case Studies in the Prelaunch Service Period," in *High Performance Computing*, ser. Lecture Notes in Computer Science, H. Jagode, H. Anzt, H. Ltaief, and P. Luszczek, Eds. Cham: Springer International Publishing, 2021, pp. 365–377.

[15] B. Schwaller, N. Tucker, T. Tucker, B. Allan, and J. Brandt, "HPC System Data Pipeline to Enable Meaningful Insights through Analysis-Driven Visualizations," in *2020 IEEE International Conference on Cluster Computing (CLUSTER)*, Sep. 2020, pp. 433–441.

[16] F. Beneventi, A. Bartolini, C. Cavazzoni, and L. Benini, "Continuous learning of HPC infrastructure models using big data analytics and in-memory processing tools," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2017, pp. 1038–1043.

[17] N. Sukhija, E. Bautista, D. Butz, and C. Whitney, "Towards Anomaly Detection for Monitoring Power Consumption in HPC Facilities," in *Proceedings of the 14th International Conference on Management of Digital EcoSystems*, ser. MEDES '22. Association for Computing Machinery, Dec. 2022, pp. 1–8.

[18] M. Molan, A. Borghesi, L. Benini, and A. Bartolini, "Semi-supervised anomaly detection on a Tier-0 HPC system," in *Proceedings of the 19th ACM International Conference on Computing Frontiers (CF)*, 2022, pp. 203–204.

[19] N. Sukhija, A. Gessinger, and E. Bautista, "Towards a Predictive Framework for Power Consumption of Jobs in HPC Facilities," in *Proceedings of the 12th International Conference on Management of Digital EcoSystems (MESDES)*, 2020, pp. 46–47.

[20] H. Shoukourian and D. Kranzlmüller, "Forecasting power-efficiency related key performance indicators for modern data centers using LSTMs," *Future Generation Computer Systems*, vol. 112, pp. 362–382, 2020.

[21] A. Bartolini, F. Beneventi, A. Borghesi, D. Cesarini, A. Libri, L. Benini, and C. Cavazzoni, "Paving the Way Toward Energy-Aware and Automated Datacentre," in *Workshop Proceedings of the 48th International Conference on Parallel Processing (ICPP Workshops)*, 2019, pp. 1–8.

[22] J. J. Dongarra, H. W. Meuer, and E. Strohmaier. Top500. [Online]. Available: https://www.top500.org/

[23] A. Bartolini, A. Borghesi, A. Libri, F. Beneventi, D. Gregori, S. Tinti, C. Gianfreda, and P. Altoè, "The davide big-data-powered fine-grain power and performance monitoring support," in *Proceedings of the 15th ACM International Conference on Computing Frontiers (CF)*, 2018, pp. 303–308.

[24] J. Thaler, W. Shin, S. Roberts, J. Rogers, and T. Rosedahl, "Hybrid approach to hpc cluster telemetry and hardware log analytics," in *IEEE High Performance Extreme Computing Conference (HPEC)*, 2020.

[25] W. Shin, V. Oles, A. M. Karimi, J. A. Ellis, and F. Wang, "Revealing power, energy and thermal dynamics of a 200PF pre-exascale supercomputer," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, Nov. 2021, pp. 1–14.

[26] S. Snyder, P. Carns, K. Harms, R. Ross, G. K. Lockwood, and N. J. Wright, "Modular HPC I/O Characterization with Darshan," in *2016 5th Workshop on Extreme-Scale Programming Tools (ESPT)*, 2016, pp. 9–17.

[27] (2024) What is a Medallion Architecture. [Online]. Available: https://www.databricks.com/glossary/medallion-architecture

[28] Constellation. [Online]. Available: https://doi.ccs.ornl.gov

[29] S. S. Vazhkudai, J. Harney, R. Gunasekaran, D. Stansberry, S.-H. Lim, T. Barron, A. Nash, and A. Ramanathan, "Constellation: A science graph network for scalable data and knowledge discovery in extreme-scale scientific collaborations," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 3052–3061.

[30] Apache Kafka. [Online]. Available: https://kafka.apache.org/

[31] ElasticSearch. [Online]. Available: https://www.elastic.co/elasticsearch/

[32] Apache Druid. [Online]. Available: https://druid.apache.org

[33] Apache Parquet. [Online]. Available: https://parquet.apache.org

[34] MinIO. [Online]. Available: https://min.io/

[35] Apache Spark. [Online]. Available: https://spark.apache.org

[36] M. Armbrust, T. Das, J. Torres, B. Yavuz, S. Zhu, R. Xin, A. Ghodsi, I. Stoica, and M. Zaharia, "Structured Streaming: A Declarative API for Real-Time Applications in Apache Spark," in *Proceedings of the 2018 International Conference on Management of Data (SIGMOD)*, 2018, pp. 601–613.

[37] (2024) Slate. [Online]. Available: https://docs.olcf.ornl.gov/services_and_applications/slate/index.html

[38] Kubernetes. [Online]. Available: https://kubernetes.io

[39] (2024) OpenShift. [Online]. Available: https://www.redhat.com/en/technologies/cloud-computing/openshift

[40] S. Peisert, T. E. Potok, and T. Jones, "ASCR Cybersecurity for Scientific Computing Integrity - Research Pathways and Ideas Workshop," Sep. 2015. [Online]. Available: https://escholarship.org/uc/item/5j00n7h2

[41] S. Peisert, "Security in high-performance computing environments," *Communications of the ACM*, vol. 60, no. 9, pp. 72–80, Aug. 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3096742

[42] Data Version Control. [Online]. Available: https://dvc.org

[43] Gitlab. [Online]. Available: https://about.gitlab.com

[44] MLFlow. [Online]. Available: https://mlflow.org

[45] A. M. Karimi, N. S. Sattar, W. Shin, and F. Wang, "Power profile monitoring and tracking evolution of system-wide hpc workloads," in *IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*, 2024, pp. 93–104.

[46] W. Brewer, M. Maiterth, V. Kumar, R. Wojda, S. Bouknight, J. Hines, W. Shin, S. Greenwood, D. Grant, W. Williams, and F. Wang, "A digital twin framework for liquid-cooled supercomputers as demonstrated at exascale," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2024.

[47] Institutional review board (irb). [Online]. Available: https://www.orau.org/orsirb/index.html

[48] W. Shin, J. A. Ellis, A. M. Karimi, V. Oles, S. Dash, and F. Wang, "Long term per-component power and thermal measurements of the olcf summit system," 2022. [Online]. Available: https://doi.org/10.13139/OLCF/1861393

[49] W. Shin, V. Oles, A. Schmedding, G. Ostrouchov, E. Smirni, C. Engelmann, and F. Wang, "Olcf summit supercomputer gpu snapshots during double-bit errors and normal operations," 2023. [Online]. Available: https://doi.org/10.13139/OLCF/1970187

[50] A. M. Karimi, B. Xie, A. K. Paul, and F. Wang, "April 2020 darshan counters from the summit supercomputer," 2023. [Online]. Available: https://doi.org/10.13139/OLCF/1865904

[51] A. M. Karimi, A. Khan, S. Oral, and C. Zimmer, "Summit darshan archival dataset," 2024. [Online]. Available: https://doi.org/10.13139/OLCF/2305496

[52] S. Atchley and M. Maiterth, "Olcf frontier supercomputer 2023-04-29 hpl power data used for top500/green500 submission," 2023. [Online]. Available: https://doi.org/10.13139/OLCF/1975494

[53] W. L. Miller, D. Bard, A. Boehnlein, K. Fagnan, C. Guok, E. Lançon, S. J. Ramprakash, M. Shankar, N. Schwarz, and B. L. Brown, "Integrated Research Infrastructure Architecture Blueprint Activity (Final Report 2023)," US Department of Energy (USDOE), Washington, DC (United States). Office of Science; Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA (United States), Tech. Rep., Jul. 2023. [Online]. Available: https://www.osti.gov/biblio/1984466