A Data Platform for Machine Learning and Data Science Applications At Large Experimental Physics Facilities

Dr. Christopher K. Allen, Senior Research Engineer
Osprey Distributed Control Systems, Inc.
304 Blue Heron Court, Ocean City, MD 21842-2452
(865)696-0670: allenck@ospreydcs.com

RE: United State Department of Energy Office of Science, Fusion Energy Sciences
DE-FOA-0002905: Topic 7, Fusion Data Machine Learning Platform

Point of Contact: Bob Dalesio: (443)834-37775: bdalesio@ospreydcs.com

PROJECT NARRATIVE

1 BACKGROUND

Osprey Distributed Control Systems initiated development of a *Machine Learning Data Platform* (MLDP) for the high-speed storage and retrieval of heterogeneous, time-correlated data available from large particle accelerator facilities and other large experimental physics facilities. It is intended as a "data-science ready" host platform for building artificial intelligence, machine learning, and general data science applications for the diagnosis, modeling, control, and optimization of such facilities. There are two primary functions of the platform, 1) high-speed data acquisition and storage of time-correlated, time-series, heterogeneous data, and 2) broad query capabilities of the archived data by data scientists and data science applications. The MLDP consists of two separate systems, the *Aggregator* and the *Datastore*. The first function of the platform is partially realized with the *Aggregator* system, which is deployed within the facility control system. It provides high-speed, real-time acquisition of online data. The remaining functionality is realized with the *Datastore*. It is a *standalone system* for high-speed archiving of acquired data and rapid data retrieval in broad forms suitable for data science applications. In addition, a standalone *Web Application* was developed as a companion to the Datastore, allowing independent inspection and interaction with the data archive using a standard internet web browser.

The MLDP was originally designed for deployment within the Experimental Physic and Industrial Control System (EPICS) popular amongst large charged-particle accelerator facilities [1]. However, during development, versality of the Datastore as a standalone system was recognized and pursued. The Datastore has an independent network communications protocol and multiple Applications Programming Interface (API) libraries requiring *no EPICS tools or systems*. Consequently, its data archiving and management capabilities can be utilized *at any facility*. Data ingestion by the Datastore has been confirmed on two separate data simulators, one emulating the Advanced Light Source (ALS) facility at Stanford Linear Accelerator, and one emulating the Materials Plasma Exposure eXperiment (MPEX) facility at Oak Ridge National Laboratory. Once data is archived, the broad data-science search and query operations of the Datastore are available to any data science application.

The Web Application was initially built as a development tool for archive inspection. The utility of this tool became immediately apparent as it provides *universal access* to the data archive. Data scientists and researchers can inspect and interact with a common archive from any remote location using a standard internet web browser.

2 Project Objectives

The project objective is to create a data platform providing universal accessibility to experimental and/or operations data collected from *any large experimental physics facility*. Archived data can be inspected, annotated, and downloaded from any remote location using a standard internet web browser. Most important, the platform is to be "data-science ready," providing query services and APIs designed for artificial intelligence, machine learning, and general data science applications.

2.1 Technical Approach

The Datastore and companion Web Application are to be formalized into a commercial-grade product deployable at any large experimental physics facility. Data from large physics experiments, or from the facilities themselves, can be archived, managed, and universally accessed with an internet web browser, or programmatically through an API library.

2.2 Status

Initial funding for the MLDP was obtained through a Small Business Innovative Research (SBIR) research grant sponsored by the United States Department of Energy (DOE), Office of High Energy Physics [2]. Extensive evaluations of the resulting prototypes were recently performed [3]. The status of the Datastore system and Web Application may be summarized as follows:

- Data from multiple disparate data sources may be simultaneous ingested and archived.
- The archive stores time-series data of type scalar, array, data structure, image, and raw data.
- Metadata associated with time-series data is stored concurrently and used for query operations.
- A consistent query service and API is offered to data scientists and data science applications.
- Universal access to all archived data is available through an internet web browser.

The basic architecture of the Datastore prototype is shown in Figure 1. The core is composed of three primary systems, the Data Archive, the Ingestion Service, and the Query Service, the latter two running as independent services on the host platform. External API libraries expose the ingestion service to data providers and the query service to data science applications. The data archive is realized by three sub-components, a MongoDB NoSQL database storing metadata, an InfluxDB time-series database, and the host file system, the latter two storing time-series data. Network communications with the core are managed with an independent remote procedure call framework based upon gRPC. Not shown in the diagram is the Web Application which has a separate host platform maintaining network connection directly to the Query Service using the gRPC protocol.

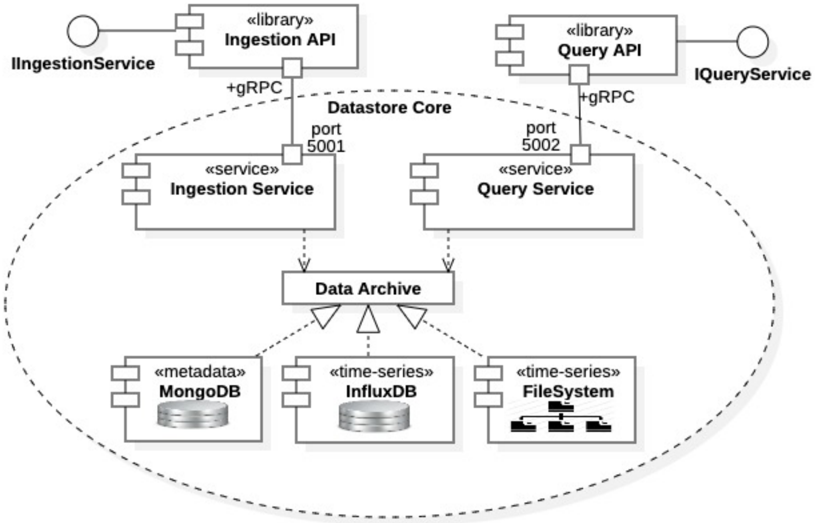

Figure 1: Datastore architecture

3 Proposed Research and Development Effort

3.1 Performance Goals - Data Transport and Archiving

The ingestion performance goal is to concurrently archive 4,000 signals at a sampling rate of 1 kHz; this goal requires a data rate of 100 Mbytes/second. Maximum data rates observed under continuous, sustained data ingestion ranged from 0.2 to 7 Mbytes/second. Thus, a project objective is to *increase ingestion performance by at least a factor 100*, ideally a factor 200.

The query performance goal is to allow data science applications to control experiment and facility operations. Thus, it is desirable to have query service performance on par with ingestion data rates, 100 Mbytes/second. The maximum data rates for the query service currently range between 2 and 2.7 Mbytes/second. Thus, *query performance must be increased by a factor 50*.

The performance criteria will require both implementation upgrades and overall archive redesign.

3.2 Archive Annotation

The ability for data scientists and data science applications to "back annotate" the data archive is a development objective. Data scientists should have the ability to inspect the archive and add notes, data associations, and calculations obtained from data science computations *post ingestion*. These annotations would then be available within the archive for later data science applications.

3.3 Web Application

The Web Application is crucial component of the project, providing universal access to the data archive. The current prototype can inspect and display archive data and metadata, it is functional but not complete. The following is a list of additional project development efforts:

- Provide archive annotation capabilities as described above (at least a viable subset).
- Provide standard data science operations (e.g., visualization, statistics, fitting, etc.)
- Identify and export specific data sets in standard formats (e.g., CSV, Excel, NumPy, etc.)

3.4 Data Science Applications

Research efforts include expansion of data science use cases in the following areas:

- Advanced tools for accessing and analyzing data sets *within the ingestion stream*.
- Creation of training data sets from archive data and online ingestion streams.
- Machine learning algorithm "plugins" for archive and ingestion stream data.

Selected data sets can be processed and culled to produce online training sets for continuous learning applications. This supports dynamic physics experiments with varying parameters where adaptive control and/or machine learning algorithms could be used in stabilization, alignment, resolution, and targeting. The feature can also be applied to facilities operations for advanced online control and optimization based upon current operations parameters. Of particular interest is the incorporation of exceptional algorithms as dedicated *plugins* for the data platform. If an algorithm is identified as exceptional it would be packaged as a plugin component for the ingestion service; it would perform fast first-level analysis making results available during ingestion. The algorithms could be packaged fully, or in parts, depending upon the complexity. For example, algorithms identifying clusters or statistics in data would likely be fully realized plugins whereas neural network training algorithms would likely be partially realized as plugins producing training data sets. Additionally, the plugins can produce metadata and other annotations for data sets that can be monitored and searched during ingestion.

RESEARCH TEAM

Osprey DCS is a small business providing state-of-the-art products and services in the areas of high-speed acquisition, large-scale control systems, networking, and data systems. We have over 20 staff members with diverse backgrounds in the areas of applied mathematics, physics, electrical engineering, computer science, data sciences, data management, and networking. Our company is uniquely suited for the design and implementation of the described data platform, having full expertise in data systems from front-end acquisition to back-end storage, archiving, and retrieval.

Dr. Christopher K. Allen, Principal Investigator, Senior Research Engineer

Mr. Bob Dalesio, Program Management, Senior Systems Architect

Mr. Craig McChesney, Senior Data Systems Engineer

Mr. Michael Davidsaver, Senior Engineer

Osprey Staff: Junior Computer Scientist

BUDGET AND BREAKDOWN

Preliminary effort and tasking breakdown with the corresponding budget estimates are provided.

Task	**Staff**	**Year 1**	**Year 2**	**Year3**	**Total**
Performance	Davidsaver (.25FTE)	76,500	76,500	76,500	229,500
Archive Redesign Annotations	McChesney (.75FTE)	229,500	229,500	229,500	688,500
Web Application	Junior CmpSci. (.75FTE)	128,250	128,250	128,250	384,750
Use Cases	Dalesio (.25FTE)	81,000	81,000	81,000	243,000
AI/ML Apps/API	Allen (.50FTE)	162,000	162,000	162,000	486,000
	Total	677,250	677,250	677,250	2,031,759

REFERENCES

[1] "Experimental Physics and Industrial Control System," Argonne National Laboratory, 12 October 2021. [Online]. Available: https://epics.anl.gov/sites.php .

[2] Osprey DCS, "A Data Science and Machine Learning Platform Supporting Large Particle Accelerator Control and Diagnostics Applications," United States Department of Energy, Office of High Energy Physics SBIR Grant #DE-SC0022583, 2022.

[3] C. K. Allen, B. Dalesio, G. McIntyre, C. McChesney and M. Davidsaver, "Machine Learning Data Platform, TM-01-2032," Osprey DCS, Ocean City, MD, 2023.