Contribution ID: **136** Type: **Oral**

MDSplusML - Optimizations for data access to facilitate machine learning pipelines

Thursday, 18 July 2024 13:40 (20 minutes)

The MDSplus[1][2] data management system is widely used in the magnetic fusion energy research community for data storage, management, and remote access. The system provides data access through a vector based, interpreter API. It was developed and optimized for rapid single shot analyses. Machine Learning applications require data from large numbers of shots and potentially from different experimental devices. We are developing tools to enable the rapid retrieval of limited sets of data from large numbers of shots. The system will cache the requested quantities in a data warehouse overnight, and be able to quickly provide them as inputs to machine learning tasks. The cache will eventually be both transparent and extensible. At this time, various caching mechanisms are being tested and benchmarked using the queries for approximately 100 quantities that are typically used by disruption-warning ML workflows. The performance of various caching schemes varies greatly depending on the environment they are deployed in. We provide comparisons of the performance of native MDSplus, HSDS[3] cache, and mongodb[4] cache in various environments. The end goal is to provide fast data access to commonly queried quantities regardless of the environment.

[1] Stillerman, J. A., et al. "MDSplus data acquisition system." Review of Scientific Instruments 68.1 (1997): 939-942.
[2] "MDSplus data system,"MIT Plasma Science and Fusion Center, April 2024, https://mdsplus.org/index.php/Introduction
[3] "Highly Scalable Data Service (HSDS)", The HDF Group, https://www.hdfgroup.org/solutions/highly-scalable-data-service-hsds/
[4] "MongoDB", MongoDB, Inc., April 2024, https://www.mongodb.com/

Speaker's Affiliation

MIT Plasma Science and Fusion Center, Cambridge

Member State or IGO

United States of America

Primary author: STILLERMAN, Joshua (MIT Plasma Science and Fusion Center)

Co-authors: Mr JELENAK, Aleksandar (HDF Group); REA, Cristina (Massachusetts Institute of Technology); TREVISAN, Gregorio (MIT Plasma Science and Fusion Center); READEY, John (HDF Group); WINKEL, Mark (MIT Plasma Science and Fusion Center); LANE-WALSH, Stephen (MIT)

Presenter: STILLERMAN, Joshua (MIT Plasma Science and Fusion Center)

Session Classification: Machine Learning

Track Classification: Machine Learning