Hewlett Packard Enterprise

Hewlett Packard Labs

Fusion's Future – Part 3 of 3: Success Depends on Developing Physics- and Data-Driven Knowledge

Author: Martin Foltin

At its core, the success of thermonuclear fusion as a technology to produce clean energy at low cost will depend on how skillfully scientists can harness the power of not only huge, expensive fusion reactors but also of data itself.

While this may seem simplistic, it's true. AI applications of all kinds require large volumes of data. There's already been plenty of fusion data collected over many decades – from fusion experiments and from the increasing volume of simulation data with the increased performance of HPC supercomputers. The research community needs a way to manage that data to be able to carefully select the most useful data to use to build AI-driven fusion models.

Labs is developing a framework that makes it easier to discover and use data that is suitable for building fusion models, and to manage these models. Applying this framework to fusion experiments will help large numbers of commercial and academic research teams across the globe advance the state of the art. As they refine the experiments that generate fusion breakthroughs, scientists need a platform that enables them to collaborate and to bootstrap new research from the results of the existing research. Ultimately, the new discoveries in fusion science will lead to the development of models that control plasma reactors. Maintaining and safely updating these models after new discoveries will be another challenge.

How does this work? The previous two articles in this series reflect the market challenge fusion faces and the ways AI models are helping to make more efficient lasers to gather more insights about fusion plasma dynamics in an inertial fusion setup. In this article, we introduce data platform enabling us to make sense of a deluge of experimental and simulation results from both inertial and magnetically confined fusion to build and periodically update fusion AI models and accelerate scientific discovery.

Fusion reactions are complicated. Hydrogen plasma is heated to extremely high temperatures (about ten times the temperature of the core of the sun) enabling high-energy collisions between two isotopes of hydrogen to form a heavier atom, releasing huge amounts of energy in the process. At these high temperatures, a number of instabilities can form in the plasma. To generate more energy, the experiments need to limit the instabilities in the system.

Defining and Tracking Instabilities

These instabilities are difficult to explain and to track. They depend on a number of variables that are statistical in nature. That makes it difficult to just do a few theoretically driven descriptions of plasma dynamics. It requires volumes of data to assimilate and compare – and researchers need a way to search data and find it efficiently.

To build AI models requires several steps. First, model builders need to find the right features in the data and build the model from these features. Our team is building the infrastructure to enable the scientists to build these models.

Capturing workflows that operate on the data allows to leverage from prior research in a reproducible manner. It extends the concept of fairness of data in fusion research. It makes data findable, accessible, and interpretable for workflows. It gives researchers the ability to make better use of not only the data but also the preprocessing algorithms that operate on that data to translate them to a more useful form for building AI models.

It also advances more than just individual models. One model feeds another model in simulation. To make the most of these complex workflows that control multiple stages, you need to be able to trace back the provenance of these models to what data was used to build these models. These models aren't stationary; they have their own life cycle. They're updated over time as new scientific knowledge comes out. Scientists need to keep track of what data has been used to build the initial model, what data has been removed from the initial data set, and what's been added, and build a new version of that model.

Building a Data Platform

Our goals are not just to control fusion experiments. Our fusion work aims to enable new discoveries and speed up the pace of new scientific discovery. Understanding the correlations between instabilities and the plasma state is data-intensive. We're trying to help scientists discover relative data to build models quicker and also to reuse existing workflows that someone already developed in academia to study specific effects.

Labs is building the infrastructure that enables data discovery and the use of workflows that are of different complexities and part of different domains. Central to this effort is the development of a Fusion Data Platform. This consists of a Common Metadata Framework that manages the metadata and the lineages of data. Having the framework in hand enables scientists to search not only for data but also for pipelines, helping them convert the data to a better form that is more usable for building AI models.

Labs has developed the infrastructure that captures these pipelines even though the pipelines receive contributions from different users across multiple sites. It enables them to reuse these workflows in the future by other institutions on other sites.

So far, Labs has developed the concept for a federated metadata framework. The framework enables scientists to record the metadata from distributed pipelines and workflows and share them with others. This way, other teams can ensure they're working off accurate models and data in real-time.

Enabling Data Discovery

The goal is to enable data discovery based just on the metadata. Moving data between different institutions is very expensive and time-consuming. So, rather than move data to assess whether it's the right data, and moving other data from place to place, our job is to enable scientists to gauge the importance of the data based solely on the metadata itself.

The next step will be to integrate the Common Metadata Framework into a larger framework managed by the Open Science Data Federation. This federation is developed by San Diego State University in cooperation with the National Science Foundation. It enables caching copies of data at different geographic locations for faster access by various research teams. By helping to identify the most useful data in context of different research problems, the Common Metadata Framework will among other things reduce the volume of data that scientists need to transfer from these caches.

Having a data framework to govern the process helps to advance scientific research. Data can be used by multiple teams or by one team with different tasks at different times. The Common Metadata Framework enables teams to track primary and derived data, models, parameters, and metrics from these workflows. This allows them to make more efficient use of existing research and save time by reusing pipelines that worked in the past.

The Data Fusion Platform is a three-year project. The first year focuses on the core infrastructure. The second year will focus on continuous learning support – managing update of models after new discoveries while avoiding forgetting relevant information from previous update. This includes the delivery of a reference implementation that provides code for pipeline stages that enable this continuous learning. Third, Labs will provide data discovery and intelligence. This may include algorithms already being used in different Labs projects to enable more efficient discovery of the relevant data on a given problem.

Conclusion

Nuclear fusion is a promising technology that could provide some answers to the world's future energy needs. But there is still a long path ahead before fusion is ready to contribute to the power grid.

With these new techniques, the time to market can be reduced. Scientists can combine physics-driven and data-driven knowledge more efficiently to make fusion more economically competitive. To do that, researchers need new ways to search for data, extract the right information from data, and transfer their knowledge to build AI models. Without data-enabling infrastructures, researchers won't be able to build AI models efficiently and a lot of knowledge would go unused.

The road map calls for reactors to be designed by the early 2030s and delivered by the middle of the decade. To get there, research needs to be accelerated, starting now.

Learn more at

hpe.com/us/en/hewlett-packard-labs

Hewlett Packard Enterprise

a00138213enw