A Scientific Cloud Computing Platform for Ingestion and Processing of SDO Data

Manuel
Indaco
Auburn University, Auburn (AL), Unites States, (Presentation by Daniel Gass)
Daniel Gass, University of Central Lancashire, Preston, United Kingdom
William Fawcett, University of Cambridge, Cambridge, United Kingdom
Richard Anthony Galvez, DataTalk AI, Dallas (TX), United States
Andrés Muñoz-Jaramillo, Southwest Research Institute Boulder, Boulder (CO), United States
Paul James Wright, Dublin Institute for Advanced Studies, Dublin, Ireland

Oral

The SDO mission has been collecting solar data for the past 13 years, producing a large dataset measured in the order of tens of petabytes. Such an immense dataset now contains information beyond that of an entire solar cycle, making it extremely valuable to the heliophysics community, becoming invaluable for critical tasks such as space weather analysis and forecasting. However, getting access to all of this data can be daunting. In the attempt to overcome the challenges associated with the management of this incredible amount of data, we successfully created a scientific computing platform, which we are now open-sourcing.

There are several reasons why access to SDO data is difficult, but they may stem from a lack of existing data infrastructure to allow researchers easy access to the dataset. As part of the 2023 FDL-X Helio challenge, we have developed a data pipeline to ingest and transform this data into a readily available, and complete data product. Using Google Cloud Platform (GCP), we host the entire 13 years of AIA and HMI data with a 6 and 12 minute cadence respectively, in 512x512 resolution. We also calibrate the AIA data up to level 1.5 and the HMI images presented as (x, y, z) componentes of the magnetic field. Furthermore, the data has been curated so that the solar disk appears the same size in each image, making the data machine-learning ready.

We will present both our data infrastructure, and the data product “SDOMLv2”, as well as discuss how we egressed the 13 years of data efficiently from its storage in JSOC. We will also discuss how our approach is generalizable and may be adopted to other scientific domains.

This work has been enabled by FDL-X (fdlxhelio.org); a derivative of Frontier Development Lab (FDL.ai); as a public/private partnership between NASA, Trillium Technologies and commercial AI partners Google Cloud and Nvidia.

Presentation file
wednesday-indaco-manuel.pdf
YouTube link
View recording
Meeting homepage
4th Eddy Cross-Disciplinary Symposium
Download to PDF