S2-D2: Securing Self Describing Data, Formats, and Libraries

Keegan Sanchez, Suren Byna, Zhiqiang Lin (The Ohio State University)
Glenn Song, Scot Breitenfeld (The HDF Group)
External Collaborators: Quincey Koziol (NVIDIA), Dana Robinson (AMD), and David Mattson (AWS)

https://idtlab.github.io/research/projects/s2d2

Introduction

- Data Management Libraries (DMLs), such as HDF5, netCDF, ADIOS2, and Zarr are critical components in scientific research.
- Several of these have been designed and developed decades ago, **without security, safety, and privacy (SSP)** as concerns**.**

1. Understanding Security for DMLs

Why: To understand how to better secure DMLs and their unique characteristics

What: Explore existing threat modeling techniques, develop a solution tailored to DMLs

2. Securing the File Formats

Why: Complex formats can lead to vulnerabilities. Data and metadata must be secured.

What: Test existing fuzzing tools, explore custom solutions, introduce anti-tampering and encryption tools.

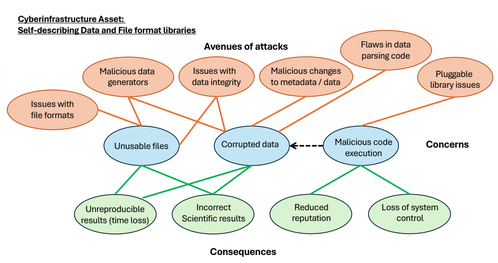
3. Securing the Software Libraries

Why: Prevent attacks being introduced by DMLs

What: Utilize fuzzing results to patch bugs. Test solutions to vulnerabilities introduced in plugins.

CASSE: A threat model for Data Management Libraries

- Problems with existing models
  - Not specific to DMLs and do not provide context.
  - Do not model the data and metadata
- Solution
  - Data modeling method
  - Three-part attack categorization
- Presented at the S-HPC ’25 workshop at SC 25, in progress journal paper

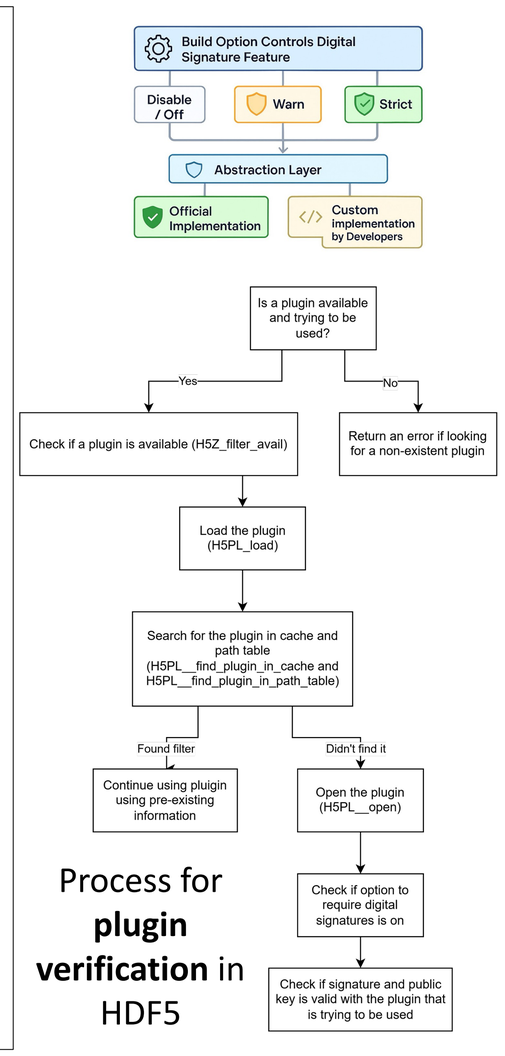


The **Three-Part** Attack Categories, split into **Source**, **Method** and **Target**

Source	Method	Target
Data	Modification	**C**ore Library
Library	Poisoning	**A**pplication
		Storage
		System
		External Library

Securing HDF5 Plugins with Digital Signatures

- DMLs use “plugins” to allow developers to extend or introduce new capabilities.
- Attackers can exploit these plugins as an entry point to attack DML based system
- A solution for HDF5
  - Verification of Plugin Authenticity
  - Detection of Tampering
  - Enhanced Trust using digital signatures
  - Enables users to confidently use plugins from reliable sources
- Presented at the S-HPC ’25 workshop at SC 25.



Process for **plugin verification** in HDF5

Ongoing Work and Other Progress

- Fuzzing DMLs – HDF5
  - Analyzed application of existing fuzzing software (AFL++)
  - Exploring solutions to improve fuzzing performance past the baseline; using LLMs for processing the fuzzing outputs.
- Encryption
  - Exploring solutions for highly parallel and configurable encryption
  - Optimized parallel encryption implementations for DMLs
- AI Readiness of Data
  - Evaluation of privacy leakage and security to assess readiness of data for AI applications
  - Developing methods to fix AI-readiness issues using AI

Next Steps

- Synthesize the tools and approaches into a security framework
  - Guide users in the application of security enhancing technologies (encryption, plugin signing, authentications)
  - Balance performance overheads with the security needs of a user.
  - Explore methods for improving performance, such as asynchronous execution.