D3 CENTER

Digital design, Datability, Decision intelligence

Cybermedia Center

Osaka University, Japan

BOOTH 4049

Provenance Recording System / Metadata Management for Experimental Data

Provenance Recording System

Background

- Ensuring reproducibility is an important requirement for Open Science and research integrity.
- To make research reproducible, a researcher must adequately record a detailed provenance, which documents the process of generating the research results.

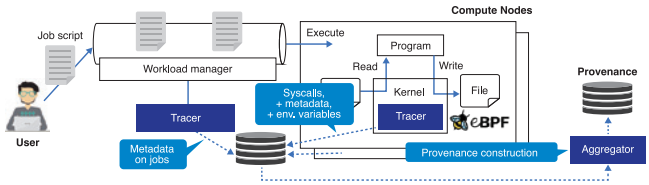
Goal

- Implementation of a low-cost mechanism for recording the provenance of data generated within computer systems, especially HPC systems.
- Requirements
  1. Provide sufficient information to reproduce and verify the computer experiment from user's perspective.
  2. Ensure transparency by requiring no modifications to users' programs.
  3. Minimize performance impacts on the applications.

Architecture

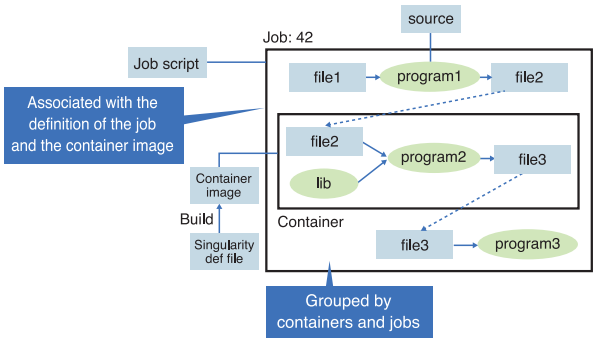
The system constructs the provenance transparently by utilizing eBPF to capture invocations of system calls in Linux kernel and gathering information about jobs from a workload manager.

- eBPF is a framework that enables attaching programs in Linux kernel.
- The Tracer identifies executed programs and files accessed by the programs by capturing system calls such as `execve()` and `open()`.
- The Aggregator constructs the provenance by merging child processes with their parent, grouping processes by jobs and containers, and associating the job script and the container image.



Constructed Provenance

Constructed provenance documents the source-to-program relationships, the job script and the container image that need to be referred to for the reproduction and the verification.



Metadata Registration and Management System for Experimental Scientists

Background

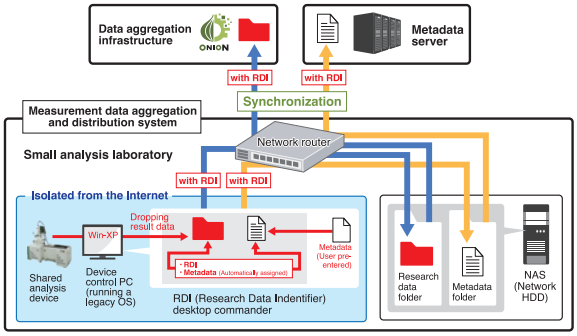
- The Core Facility Center developed a measurement data aggregation and distribution system that enables researchers to collect measurement results in a secure manner from analytical devices that are isolated from the Internet.
- The system is connected to the data aggregation infrastructure, ONION (Osaka university Next-generation Infrastructure for Open research and open InnovatioN) of the D3 Center via the S3 protocol.
- The problems in the current system are:
  - The absence of metadata and identifiers prevents the utilization of the data.
  - It is usual to be generated large amount of failed measurement results.
  - Researchers must assign metadata manually.
- The objective is to develop a research data/metadata aggregation and management infrastructure that enables researchers to assign and manage metadata with minimal additional burden.

Conceptual image

- It would be desirable for a system to be able to automatically assign identifiers and the minimum necessary metadata, and only assign detailed metadata for expected results.

Requirements

1. The system should target all research data, which must be managed and traceable.
2. The system should facilitate the management and tracking of research data by the researcher.
3. The system should be designed with data utilization in the context of open research data.
4. The system should never place a burden on experimental scientists, such as assigning metadata to every generated measurement data.
5. The system should employ a method of metadata collection and aggregation in alignment with the research practices of experimental scientists.



Contact : sc24@ais.cmc.osaka-u.ac.jp https://www.nri.cmc.osaka-u.ac.jp/