Provenance Recording System for Research Data Management

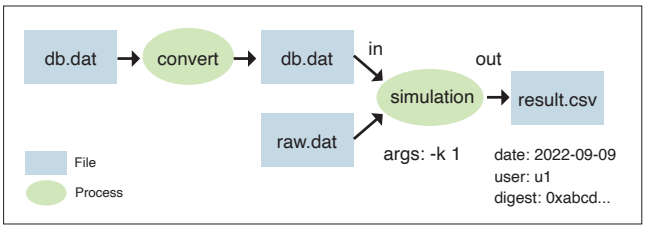
Background

- Growing importance of research data management (RDM)
  - To ensure **reproducibility** (transparency): Preserving data that provide evidence of research results
  - To improve **reusability**: Promoting the global sharing of knowledge and increasing research efficiency

- Provenance, which identifies the input data and the process used to obtain data, should be secured for reproducibility and reusability

- HPC systems generate data through simulations and experiments, but there is no established method to manage the provenance of the data

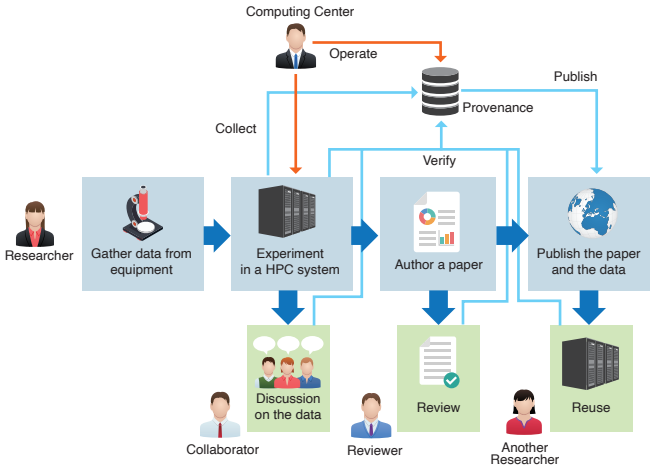
- A system that implements RDM on HPC systems is needed



Example of a provenance

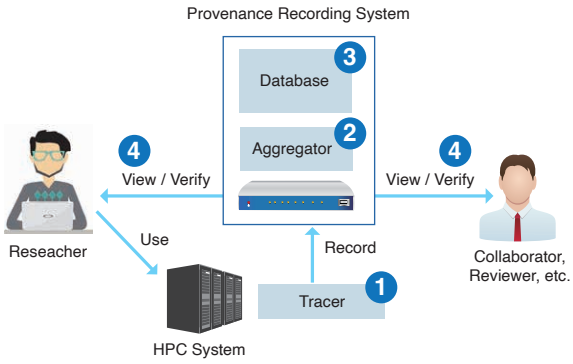
Requirements for Provenance Recording System

- Automatically record the provenance and the metadata (date/user created, etc.) of a file generated in a HPC system
- Support a typical HPC environment: workload manager (Slurm), MPI, etc.
- Minimalize impacts on performance and user's operations
- Secure the records not to be falsified
- Provide interfaces to verify that a file has not been fabricated/falsified

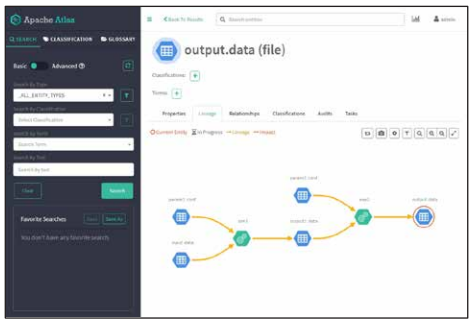


Use cases

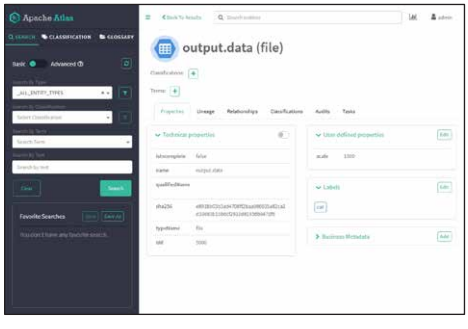
Prototype



1. Tracer captures system call invocations (exec(), open(), write(), etc.) of a user program. BPF, a low overhead observability scheme in Linux kernel is used for the capture. Tracer also captures metadata (date created, SHA-256, etc.). No modifications in the user program and operations are required.
2. Aggregator builds a provenance of files from the history of the system call invocation: a file read/written by a process is an input/output of the process in the provenance. Parallel processes by MPI are aggregated.
3. The provenance and the metadata are stored in Apache Atlas (an open-source data catalog).
4. Find and verify the provenance and the metadata of a file (shown below).



Show the provenance of a file



Show the metadata of a file

This work was carried out in Joint Research Laboratory for Integrated Infrastructure of High Performance Computing and Data Analysis
https://www.nri.cmc.osaka-u.ac.jp/