SDCC Data Management Policies

Introduction

The US Office of Science and Technology Policy (OSTP) and the US Department of Energy (DOE) Office of Science (SC) require that all proposals submitted to the Office of Science must include a data management plan (DMP). In addition to describing what and how data generated by the proposed research will be shared and preserved, the DMP needs to identify data management resources at any facility that may be used in the course of the proposed research and any associated policies for these resources This document discusses the data management resources and policies at the Scientific Data and Computing Center (SDCC) at Brookhaven National Laboratory and can be used as a reference by Principal Investigators when creating data management plans for their specific projects.

High Level Summary

The SDCC provides a range of data management services as part of a portfolio of computing and storage services at the data center. The services differ in their availability, reliability, durability, accessibility and performance characteristics, to accommodate differing requirements for specific types of data. These services are available to authorized researchers and collaborations to help them in their pursuit of scientific discovery. This document covers characteristics and capabilities of these data management resources and policies governing their use so that researchers and collaborations can determine where they might fit in their research plans and in their DOE mandated Data Management Plan. **In general, the SDCC provides operational support for data management services that are actively funded by the specific projects that utilize the service. A project's long-term access to data management services beyond the funded period is not guaranteed without a written service-level agreement between the project and the SDCC.**

Analysis Code and Metadata

As a shared facility, the SDCC provides data management services to large numbers of independent users and collaborations. For this reason, unless otherwise negotiated with data owners, the SDCC treats all data as "opaque" data, meaning that the facility treats all data as a sequence of bytes. This means that the SDCC does not interpret or analyze the data stored at the facility. It is the responsibility of users and collaborations that own the data to maintain the necessary metadata and software to read and analyze the data stored at the facility, now and in the future. In addition to the project specific software, this includes additional support libraries and operating system versions needed to run the specific software in the future. Finally, it is the responsibility of the owners of the data to place their data in the correct facility provided data managements services, relative to the requirements of the data. Any precious data requiring special dispensation should be identified to the SDCC so that the appropriate quality of service can be provided if possible.

Data Confidentiality and Access Controls

The SDCC is a shared, open research facility intended for fundamental scientific research. Priority is given to facilitating access to data and sharing of information between users at the facility. Most data management services at the facility support moderate level of access control and confidentiality. In addition, to provide maximum convenience and flexibility, responsibility for configuring access to data has been given to individual users or collaborations that own the data. The SDCC data management resources are not designed to hold highly confidential data, personally identifiable information (PII) or data covered by HIPAA regulations. Finally, while care is taken to secure the facility, security breaches may result in unintended exposure to data given the capabilities and characteristics of the data management services provided by the facility.
Note that SDCC system administrators with "root" or "administrator" privileges have the ability to bypass all access control mechanisms. Selected vendor support personnel, under the supervision of SDCC staff, may also have the ability to bypass access control mechanisms. "root" privileges are used only under certain highly restricted situations, and typically only used to look at user's files when there is a operational problem or a security issue. In the former case, this is typically in response from owner of the file. Following are common instances where a user's file may be accessed by an SDCC administrator:

- Error, output, and job log files to determine if a job or access failure was due to user error or a system failure.
- If a user explicitly request assistance from facility administrators, via any mechanism, e.g. help ticket, direct personal email, in person, etc., the request is considered to be explicit permission to view the user's files as needed to resolve problems identified by the users.

Under ordinary circumstances, SDCC will not copy, expose, discuss, or in any other way communicate a user's data to another person. However, there are two key exceptions:

- When a user account expires or a user leaves a project, the SDCC will honor requests to change file ownership when instructed by the user or the most recent principal investigator (or designated PI proxy) of the project to which the user was associated.
- The SDCC is required to address, safeguard against, and report misuse, abuse, security violations, and criminal activities. As defined in the BNL computer use agreement, SDCC retains the right, at its discretion, to disclose any or all data files or records of network traffic to appropriate cybersecurity organizations and law enforcement officials.

Transient Data Storage Services

The SDCC maintains five types of transient data stores with different characteristics:

1. Home directories
2. Document file systems
3. Project file systems
4. Software distribution file systems
5. Object Storage
6. Cold Storage Services

7. Project dedicated data storage
8. Source code version control repositories

As mentioned previously, users and collaboration should not store highly confidential data, PII data or data covered by HIPAA regulations.

Home Directories

All SDCC users are provided with a home directory of limited size. This file system is accessible from systems at the facility including compute nodes and data transfer nodes (DTNs). Home directories are designed for source code and documents and are optimized for small to medium size files. Home directories are not designed from high bandwidth or high IOPS access to data. In addition to being backed up on a daily basis, daily snapshots of directories are also made. Directory backups are kept for 90 days, while snapshots are kept for 7 days. Note that the backup and snapshot mechanisms are designed for infrequently modified text files and cannot guaranty file consistency for frequently changing text or binary files. Home directories for inactive users (i.e. no longer BNL employees, no longer associated with sponsoring collaborations, or no longer authorized Guests at BNL) will be migrated to long term archives at the discretion of the SDCC with input from authorized collaboration representatives.

Document File Systems

Document file systems are targeted at groups of users or collaborations, typically for sharing small data, documents, or applications. They have the same characteristics and usage constraints of home directories but are not backed up nor are snapshots made, unless explicitly negotiated with the facility.

Project File Systems

Project file systems at the SDCC are targeted primarily at collaborations for high bandwidth, large block access to large files. These file systems are accessible from systems at the facility including compute nodes and data transfer nodes. Neither snapshots nor backups are made of project file systems. Some project file system, particularly but not limited to those designated as “scratch”, may have policy driven auto-deletion of files. If present, these policies are negotiated with the “owners” of the file systems.

Although designed for high performance, there is no provision for high bandwidth access to project file system from the general internet. Although the facility data transfer nodes make these file systems accessible from the internet, these DTN nodes are designed for occasional, moderate bandwidth access to project file systems. Any sustained, high bandwidth access to project file systems from outside of BNL needs to be negotiated with the SDCC.

Software Distribution File Systems

Software distribution file systems are read only, world accessible and readable file systems. These file systems are designed for worldwide, read only distribution of applications and libraries. The file systems do not provide any access controls, although there is a level of data integrity through the use of cryptographic hashes. While not actively backed up, there are independent copies of the data on

redundant systems.

Object Storage

BNL Box at the SDCC is used to provide worldwide access to object storage. The service is designed as a near line repository for “cool” data, i.e. data with limited, read mostly access. Write access to BNL Box is limited to authorized users and read access to files is controlled by the owner of the data. Data in BNL Box is not backed up, although data is stored on a storage system that is protected with RAID and erasure codes. The SDCC provides tools to help users move or copy data to more durable storage.

Cold Storage Services

In addition to the transient data storage services mentioned above, the SDCC provides limited access to data storage on magnetic tape. This service is for “cold” data, i.e. data that will be accessed infrequently, if at all. This service is tailored for large files (files > 10GB in size). Access to cold storage must be negotiated with the SDCC as the characteristics of the service provided can vary dramatically depending on access and quality of service requirements.

Project dedicated data storage

The storage services mentioned above are the services that are available to all users and collaborations. For certain special cases, the SDCC also support dedicated data management services for specific groups or collaborations. Potential services in this category are negotiated on a case by case basis and must leverage existing capabilities available at the facility.

Source Code Version Control Repositories

The SDCC provides Git source code version control repositories for our users, with a web interface supporting additional collaborative functionality. This is a low bandwidth service that is not backed up. Due to the distributed nature of Git, users are expected to have copies of their repositories available locally for data recovery purposes.