ARCHIVER Project

Technical Summary

PETRA III / EuXFEL DATA ARCHIVING

Problem Definition:

PETRA III is the worldwide most brilliant storage ring based X-ray sources for high energy photons. 22 beamlines distributed over three experimental halls are concurrently available for users. The European XFEL is a world's largest X-ray laser generating 27 000 ultrashort X-ray per second and with a brilliance that is a billion times higher than that of the best conventional X-ray radiation sources. 6 beamlines are available for user experiments.

The two facilities produce yearly about several 10s PB of raw data and this is expected to double in size every year. Currently the data is automatically transferred and stored in DESY data center in two different systems: a high-performance parallel file system (i.e.. IBM Spectrum Scale) serving as a cache and/or for data analysis and a hybrid disk/tape-based storage system managed by dCache. Always at least two copies of data are stored in the datacenter.

With an increasing demand on storage space and limited on-prem data center capabilities and flexibility, cost-effective cloud-based data archiving might be a good alternative or part of a hybrid on-prem/cloud-based solution. Storing the data or a subset of the data in a cloud will also allow to relatively easy open research data for public access.

So, the goal is to replace the currently used solutions for archiving the two copies of PETRA III /EuXFEL data with a new service developed during the ARCHIVER project.

Lifecycle - Workflow Characteristics:

1. Individual photon scientist's data archiving

The individual photon scientist (user) should be able to create an archive for his thesis, publication, or generally all data which has a personal binding to that individual scientist. There is no generic recommended solution for this problem at DESY at the moment. In some cases we create so called "virtual beamline" so that a user can inject data which follows the workflow of a real experiment by copying data to data center, making tape copies, etc.

Single archive size: average 10-100 GB.

Files in archive: average 10,000
Total archive size per user: 5 TB
Duration: 5-10 years
Ingest rates: 10-100MB/s (more is better)
Data is personal, does not require encryption - optional, nice to have

Since the data sizes are relatively small, all elements of this workflow can be done via a web browser:

A. Registration to the system. The user obtains an initial account from an administrator or applies to the service so that administrator can allow him to use his DESY account or some public account (e.g. Google, Facebook, etc.). During this stage a predefined policy is set for this account. The policy includes:
   - maximum and default lifetime of the archive,
   - maximum total archives size,
   - default quality of services such as number of copies, location (e.g. according to EU GDPR), checksum policy, and others,
   - mandatory roles: e.g. owner/manager (see workflow 3)

B. Login to the system: the user can login to the system using one of his credentials such as his/her DESY account via Edugain, a local or public account. All that will be mapped to a single identity (DN) for which the configured authorizations apply.

C. Create a new archive: the user opens a new archive (starts with an empty one), names the archive. Then, he/she ingests data by uploading via a web browser (folder-wise or individual files). The user can upload data using other sessions (opened in other browsers on other locations). After all data is uploaded, he/she can close the archive. At any time, the user can add metadata to the archive in formats such as key-value strings or simple strings. After the archive is closed, all data is considered immutable and the only allowed operations is to change/extend the metadata (this has to be recorded/tracked by the archive system) or to delete the archive.

D. Close an archive: the user will get a DOI for that archive.

E. Searching for an archive: the user can list all his archives. He/she can search archives with metadata matching the search criteria. This searches (queries) are interactive and requires immediate responses (recursive, higher detailed searches - i.e. similar steps like internet searches done)

F. Reading an archive: the user can download the whole archive, selected folders and/or files from the browser or get a download link. He/she can see checksum and can use a tool to verify the downloaded data.

G. Sharing an archive: the user can allow other users to work on his archive with read only, or read/write permissions by adding another user and a respective role to the system, either on the account level or on the individual archives. The user can get a shared link which allows to download (e.g. via https) the whole archive, individual folder, individual files.

2. Manual data archiving for an individual beamline experiment

The experiment manager should be able to archive data collected during a specific photon science experiment. This workflow mainly relate to experiments done on one of the 22 beamlines of PETRA III facility. Each experiment might be quite individual, which does not allow fully automated processing. Therefore, the data should be archived based on pre-defined automated workflows and manually by using corresponding API/CLI as the data volumes are too large to allow web browser based submissions. Metadata handling inherited from the previous workflow.

Single archive size: average 5 TB
Files in archive: average 150,000
Total archive size per beamline: 400 TB, doubles every year
Duration: 10 years
Ingest rates: 1-2GB/s
No personal data, no encryption required

The elements of the workflow are:

A. Registration to the system: the experiment manager obtains an initial account from an administrator or applies to the service so that administrator can allow him to use his DESY account or some public account (e.g. Google, Facebook, etc.). During this stage a predefined policy is set for this account. The policy includes:
   - maximum and default lifetime of the archive,
   - maximum total archives size,
   - default quality of services such as number of copies, location (e.g. according to EU GDPR), checksum policy, and others,
   - mandatory roles: e.g. owner/manager

B. Group management (delegation): The experiment manager can add other users and set their roles. These roles can be set globally or per archive.

C. Login to the system: a user can login to the system using one of his credentials such as his/her DESY account via Edugain, a local or public account. After a login, a token is obtained to use API/CLI. Obtaining a token via a non-web session would be beneficial.

D. Create a new archive: the user opens a new archive (starts with an empty one), names the archive. Then, he/she ingests data by uploading via a web browser (folder-wise or individual files). The user can upload data using other sessions (opened in other browsers on other locations). After all data is uploaded, he/she can close the archive. At any time, the user can add metadata to the archive in formats such as key-value strings or simple strings. After the archive is closed, all data is considered immutable and the only allowed operations is to change/extend the metadata (this has to be recorded/tracked by the archive system) or to delete the archive. The user should be able to perform all operations via API/CLI by using predefined workflows/setups (i.e.

archive raw and calibration data which are largely in common to all experiments). These workflows determine the source specification, data transport methods, etc.

E. Close an archive: the user will get a DOI for that archive via API/CLI.
F. Searching for an archive: the user can list all his archives via API/CLI. He/she can search archives with metadata matching the search criteria.

G. Reading an archive: the user can download the whole archive, selected folders and/or files from the browser, API/CLI or get a download link. He/she can see checksum and can use a tool to verify the downloaded data.
H. Sharing an archive: the user can allow other users to work on his archive with read only, or read/write permissions by adding another user and a respective role to the system, either on the account level or on the individual archives. The user can get a shared link which allows to download (e.g. via https) the whole archive, individual folder, individual files. This operations can be done via API/CLI.

3. 3. Integrated data archiving for large standardized beamline/facility experiments

Some of the beamline experiments are quite standard and done using the same hardware (detectors, etc), software, file format, data structure. This allows the creation of a fully automated service from data taking to data processing (Experiment As a Service). Therefore the solution should allow to include data archiving in this automated workflow so that no physical person will interact with the archive. This scenario is mainly related to the European XFEL facility (facility taking care of data archiving) and also characterized by very high data volume/ingest rates.

Single archive size: average 400 TB.
Files in archive: average 25,000
Total archive size per beamline: 10s PB, currently doubles every year
Duration: 10 years
Ingest rates: 3-10GB/s - averaged over 1-3 hours
No personal data, no encryption required

The elements of the workflow are:

A. Preparation of integration: registration to the system, role management, access token, workflow/setup.
B. Integration of the archiving service: the API/CLI provided by the archiving system should allow to integrate it in an existing workflow. This basically means triggering the archiving process as soon as data is ready to be archived. E.g. after experimental data is collected and preprocessed, the archiving system picks up from a given folder, creates an archive, sets policies, fills it with data, adds metadata, closes an archive and returns DOI. This requires the typical operations listed in previous cases (i.e. creating a new archive, searching, reading, sharing) should be automated. In addition the archive system should

be able to generate events for certain conditions (i.e. data and metadata has been accepted, stored and verified) to allow data deletion from online storage. Other conditions to generate events could be a read access, metadata changes, QOS changes such as changing data storage policy, and others.

C. In order to achieve the required data rates, RDMA based protocols and ‘third party copy’ operations should be supported.
D. In order to support archiving new versions of derived data and in general additional data, the system should be able to support additional data to be appended by inheriting configurations and metadata from the original archive object. The original data will not change, the new archive object will be shown similar to a ‘union filesystem’. This feature is required by all workflows in this deployment scenario.

Authentication and Management Functions:

For this particular deployment scenario, any type of admin activity and thus, authentication in the admin role is not required to be based on ‘non local accounts’.

Managing archive object access on the other hand, requires many supported authentication method such as X509, OpenID, eduGain and others. The archive object owner/manager has to select user credentials and use them to control access rights. The archive must be able to verify and handle all supported authentication methods. For web browser access we would need DESY account via Edugain, a local or public account to login to the web browser.

For API/CLI the use of tokens seems mandatory.

Data Storage (Archival Storage):

Instances of ‘Archival Storage’ (i.e. Tape, S3 store, etc.) should be configurable in a distributed way (onsite and/or offsite) in a horizontal (i.e. for replication) and/or vertical (tiered) configuration. Protocols to transfer data between them should be standard based, ‘firewall friendly’ and provide a decent efficiency for LAN and/or WAN transfers.

Data and Metadata Characteristics:

All data is unstructured. Their metadata have no specific format but generally keep an indexable key-value format as is a requirement with up to 1 KB of size for key and 100 KB for value. Typical are 1KB of metadata per GB of data with large variations. Value should support: binary, string, date, number. We expect no personal data in any of these datasets. Data migration or reformatting is not required in this deployment scenario.

Data Policies and archive profiles:

Site and community data policies together with ‘contracts’ between archive service responsible and the local community should fully determine all parameters characterizing the archive operation and SLAs. It will fully determine the costs for archive data operation for the duration agreed on. We expect a decent number of these ‘archive profiles’ to be defined and configurable authorizations to allow a specific subset of archive profiles to be selectable by specific users (roles).

Interface Characteristics:
As it is mentioned in the workflows above, we require, browser, API and CLI methods to interact with the archive service.

Reliability Requirements:
Service must be up 99.999999999%. Data reliability is up to the negotiated archive profiles as a tradeoff between costs and data reliability/availability/integrity.

Compliance and Verification:
No product or service compliance required. Checksumming with transparent and user readable checksums both for data and metadata is required

Cost Requirements:
Costs need to be predictive over a long time (5-10y). Overall costs (and QOS) needs to be comparable to the current in-house effort.

Initial Data Management Plan:

DMP Topic	What needs to be addressed
Data description and collection or re-use of existing data	Raw, Calibration and derived data are going to be handled in ARCHIVER. Raw and Calibration data is produced by detectors directly - other (derived) data is generated by computer based data analysis. In most cases all data is in HDF5 format and handed over to ARCHIVER for long term storage (cold data) and to generate a second copy (with respect to the primary storage used while high access rates are expected) as soon as possible.
Documentation and data quality	https://www.xfel.eu/data_privacy_policy/index_eng.html https://www.xfel.eu/users/experiment_support/policies/scientific_data_policy/index_eng.html
Storage and backup during the research process	Today, DESY is using its dCache/OSM based service to archive ‘cold data’ in a HSM like (disk + automated tape) service. That layer is going to be replaced/extended with

	ARCHIVER services using an additional copy of the data. (cleanup at the end of the project requires no further data movement).
Legal and ethical requirements, codes of conduct	The selected scientific data used in Phase II+III will not contain any personal data. Ownership and IP property is addressed in the scientific data policy (https://www.xfel.eu/sites/sites_custom/site_xfel/content/e51499/e51503/e52947/e56273/e56274/xfel_file56275/ScientificDataPolicyapprovedbyCouncilon30June2017_eng.pdf) Section 5.
Data sharing and long-term preservation	Open Data is already covered by the selected data policies (second item) and selected (subset) scientific data stored in ARCHIVER, will be candidates to verify open data handling according to the policies mentioned before. All data will be HDF5 formatted - HDF5 is provided by the HDF collaboration (https://www.hdfgroup.org). The idea to use a dedicated copy of open data residing in a public cloud, will include a seamless method to use public cloud compute resources to process that scientific data.
Data management responsibilities and resources	Responsibilities are addressed in the scientific data policy (https://www.xfel.eu/sites/sites_custom/site_xfel/content/e51499/e51503/e52947/e56273/e56274/xfel_file56275/ScientificDataPolicyapprovedbyCouncilon30June2017_eng.pdf) Section 4 covers the main responsibilities