On-Site Data Management for ANKA

Halil Pašić

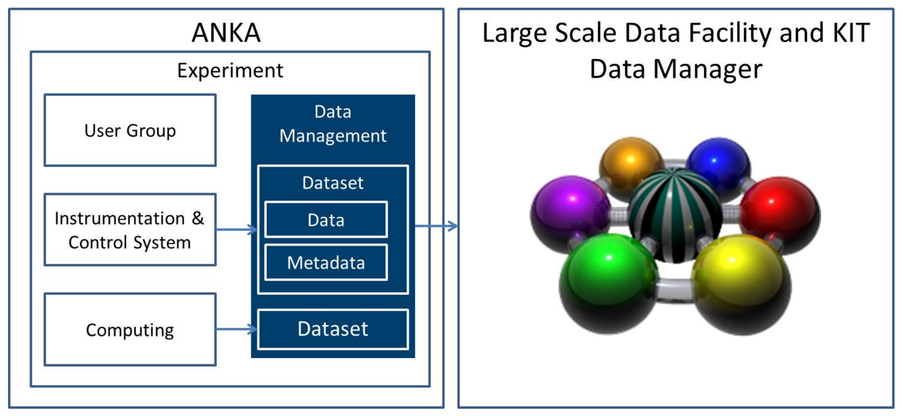
Requirements

ANKA is a synchrotron facility with 16 beamlines attached. The most important requirements of ANKA towards data management are the following:

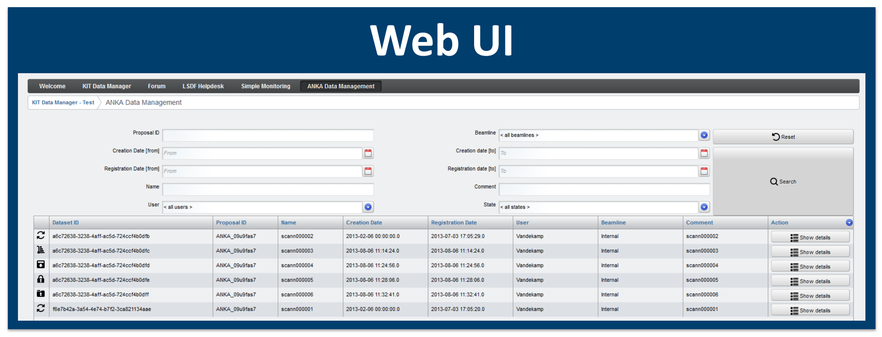
- **Automated data management:** Users do not need to be involved into technical details of data management and data policies.
- **Flexibility:** The continuous improvement of the experimental setups is an important part of ANKA research.
- **Large scale data:** ANKA produces multiple petabytes of data yearly.
- **Long living data and metadata:** Data which lead to publications needs to be archived for at least 10 years.
- **Heterogeneous data:** The beamlines support many experiments and users.

Dataset Based Data Management

- The Large Scale Data Facility and the KIT Data Manager together form a peta-scale repository, both available and suitable for ANKA data.
- The primary data is created in the beamlines. Hereby local storage resources are used.
- The data created in the beamlines organized into datasets (digital objects in repository nomenclature) from the very beginning by the on-site data management. This enables policy based data management and an easy ingest into the repository.



Software for On-Site Data Management



Administrators can

- Inspect the state of the system
- Request a state modification

DatasetDatabase: Represents the shared (and persistent) state of the distributed system.

- Holds descriptive and structural information about the datasets and the resources used by it.
- Serves as a fundament to data management decisions.
- Implemented as a transactional relational database.

DatasetDatabase

HandlerServerDB

- Data Propagation Policy
- Replicate Dataset Data
- Move Dataset Data
- Ingest Dataset

HandlerServerDB: Performs transformations on the DatasetDatabase.

- Orchestrates complex actions.
- Implements policies.

Beamline 1

WorkspaceCreator

- UUID
- Metadata Extraction
- Resource Acquisition

HandlerServerStor

- Close
- Replicate
- Delete
- Disk Free
- Check Integrity

...

Beamline n

- WorkspaceCreator
- HandlerServer

WorkspaceCreator: Initializes the data management environment for data of certain kind.

- Plug-in architecture (to accommodate experiment specific logic easily)
- Tango (CORBA) API for integration with the control system

HandlerServerStor: Interacts with the storage resources to execute tasks specified by the DatasetDatabase.

- Plug-in architecture for extensibility

Conclusions

- Dataset based data management
- Zero overhead for file IO (e.g. during measurements)