CDI HTTP-API environments at EUDAT

TTG9 - 27TH OCT 2020
The EUDAT Team in SDC
CDI: Behind the scenes

Ingestion
Unrestricted orders
Restricted orders

CDI HTTP APIs

SeaDataCloud

EUDAT
EUDAT Services – B2*

Ingestion
Unrestricted orders
Restricted orders

CDI HTTP APIs

B2STAGE

B2SAFE

B2HANDLE
Register your Research Data
B2SAFE – data management

- Built on top of iRODS, the core data technology of the EUDAT infrastructure
- Implements customizable data storage and management policies
- Allows for data replication across geographical domains
B2STAGE – data transfer

- Extensible RESTful HTTP interface
- Functionalities for data transfer between EUDAT resources and external facilities
- Built on top of B2SAFE
B2HANDLE – data identification

- Distributed service that manages Persistent Identifiers (PIDs) for data hosted onto the EUDAT network.
- Resolution service is based on the Handle System to allow for user forwarding to the current object location.
Ingestion workflow

RM (Replication Manager)

IM (Import Manager)

HTTP API

QUALITY CHECKS

CELERY ASYNC WORKER

Local Cache

B2HANDLE

PRIVATE DOCKER REGISTRY
Unrestricted ordering workflow

Order URL

HTTP API

PIDs

CELERY ASYNC WORKERS

Local Cache

SeaDataCloud

RM (Replication Manager)

IM (Import Manager)

29/04/2020
Restricted ordering workflow

SeaDataCloud
RM (Replication Manager)
IM (Import Manager)

HTTP API

ORDER URL

Local Cache

B2HANDLE

CELERY ASYNC WORKER

request

FILES

FILES
Three different environments

- Development (DEV) environment
- Test environment
- Production (PROD) environment
DEVELOPMENT environment

- Currently installed at CINECA
- Deployment of latest development version
- Only accessible to other developers (EUDAT, MARIS, IFREMER)
- Potentially untested/untrusted functionalities
- Temporary data
TEST environment

- Currently installed at **CSC**
- Deployment of latest **stable version**
- Working functionalities but **temporary data**
- Accessible to **all SDC partners**
PRODUCTION environment

- Currently installed at CSC
- Deployment of latest stable version
- Production data (4.5M files)
- Accessible to all SDC partners
CSC deployment (TEST + PROD)

- 8 VMs, 44 cores, 120gb RAM
  - B2STAGE / HTTP API + NFS (400gb) (6cores, 16gb RAM)
  - iRODS (6+3 cores, 16+4gb RAM)
  - Celery (8 workers) + Redis + Celery Backend (8 cores, 32gb RAM)
  - Rancher (3 cores, 4gb RAM)
  - Host of quality checks (8 cores, 24gb RAM)
  - B2STAGE TEST + Celery TEST + NFS (350gb) (6 cores, 16gb RAM)
  - Host for quality Checks TEST (4 cores, 8gb RAM)
  - + remote DockerHub (GrNet)
Data synchronization

- All ingested data is automatically synchronized on different EUDAT centers
- In particular from CSC to DKRZ and CINECA
Warm replication at DKRZ

- Data is synchronized at DKRZ to create a data archive used for data preservation
- The implemented workflow prevents the propagation of accidental data loss and files deletion
- Can be used for disaster recovery
- Status: operative
Stand by replication at GRNET and STFC

- Ready and update B2SAFE instance
- Can be used as temporary solution when the Host replication instance has a problem
- Status: operative
Hot replication at Cineca

- Data is synchronized at Cineca to create an online mirroring of the system
- This archive will be linked to a new production environment to offer unrestricted orders functionalities
- Will be used for services redundancy and high availability
- Status: system setup – coming in the next weeks
Data replication
PID resolution – master PID
PID resolution – Replica PID

ORDERING MASTER PID

CSC

CINECA

WORK IN PROGRESS

Master PID

File location

Master PID

File location

Master PID

File location

Replica PID

iRODS

iRODS

iRODS
Monitoring

Production (PROD) environment

Ingestion
Unrestricted orders
Restricted orders

Monitor APIs

<table>
<thead>
<tr>
<th>Action</th>
<th>CDI HTTP APIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authentication</td>
<td></td>
</tr>
<tr>
<td>Create a batch (CELERY)</td>
<td></td>
</tr>
<tr>
<td>Check if the ingestion is completed</td>
<td></td>
</tr>
<tr>
<td>Quality check</td>
<td></td>
</tr>
<tr>
<td>Delete the batch</td>
<td></td>
</tr>
</tbody>
</table>

Monitor B2SAFE

<table>
<thead>
<tr>
<th>Action</th>
<th>CDI HTTP APIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authentication</td>
<td></td>
</tr>
<tr>
<td>Upload file</td>
<td></td>
</tr>
<tr>
<td>Get File</td>
<td></td>
</tr>
<tr>
<td>Delete file</td>
<td></td>
</tr>
</tbody>
</table>

Monitor B2HANDLE

<table>
<thead>
<tr>
<th>Action</th>
<th>CDI HTTP APIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mint</td>
<td></td>
</tr>
<tr>
<td>Update</td>
<td></td>
</tr>
<tr>
<td>Resolve PID</td>
<td></td>
</tr>
<tr>
<td>Delete PID</td>
<td></td>
</tr>
</tbody>
</table>

CDI
HTTP
APIs

B2SAFE
B2HANDLE

SeaDataCloud
Conclusions

- **HTTP APIs (B2STAGE + custom SDC functionalities)**
  - Data transfer functionalities and interface of all other services
- **iRODS (B2SAFE)**
  - Data storage (replicated among all involved centers)
- **Persistent Identifiers (PIDs, B2HANDLE)**
  - Data identification
- **Celery**
  - Asynchronous operations (e.g. data ingestion and data harvesting)
- **Rancher & Docker**
  - Execution of quality checks