Onedata FaaS Workflow Engine for Archivisation and beyond

Presented by: Lukasz Dutka
IN COLLABORATION WITH

T-Systems

GWDDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)
**Bagit Files**

**Anatomy**

- **SIP.** Popular format for submission packages.
- **Metadata.** Delivers metadata information about the content.
- **Checksums.** Contains hashes of data and itself for consistency validation.
- **Deliver Data.** Delivers data for smaller collections.
- **List of Data to be fetched.** Describes the way to fetch (PULL) into the system.
- **Standard?** Unfortunately there many flavours of Bagits and several custom extensions.
NEW FEATURES
AUTOMATION ENGINE
01
**Lambda Anatomy**

- **Input Arguments**: `<Map>`
- **Mount Space as File system**: `<Oneclient>` optional
- **Output Results**: `<Map>`
- **Exceptions**: `<Map>`
- **Logs**: `<Map>`
- **Side-effects**: e.g. REST-API calls
- **Heartbeats**: For long running lambdas
- **Stores Updates**: Direct operations on stores
- **Batch Mode**: Can work with batches of input arguments to speed up the process
**Workflow Anatomy**

- **Lanes.** Iterates over Store and execute parallel boxes
- **Stores.** Input to to the workflow or produced during the workflow
- **Parallel Boxes.** Contains Lambdas which can be executed in any order
- **Lambdas.** Function which is called by mapping arguments

- Can be exported to JSON and reused by someone else
Store Anatomy

- **Persistent.** Keeps information to be iterated
- **Internal Model.** List, KV Map, Single Object, Forest Tree, Histogram for time series data
- **Strict Types.** One of: Object, File, AnyFile, Directory, String, etc.
- **Input User.** Defined before workflow execution.
- **Browsable.** User can see the current and saved status of all stores until the workflow execution is purged
Inventory Anatomy

- **Workflows.** Keep the list of workflows to be available for system users
- **Lambdas.** Keep the list of registered Lambdas
- **Members.** Access control
- **Import/Export.** Import full workflows into Inventory from JSON file
NEW FEATURES
OAIS ARCHIVISATION
02
### Datasets, AIP, DIP, Incremental Archives

**DATA**

- **My Collection**
  - **AIP**
  - **DIP**

**Files**

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Modification</th>
</tr>
</thead>
<tbody>
<tr>
<td>bag1.txt</td>
<td>58 B</td>
<td>28 Jul 2021 10:53</td>
</tr>
<tr>
<td>data</td>
<td>—</td>
<td>28 Jul 2021 10:53</td>
</tr>
<tr>
<td>manifest-md5.txt</td>
<td>5.8 KB</td>
<td>28 Jul 2021 10:53</td>
</tr>
<tr>
<td>manifest-sha1.txt</td>
<td>6.2 KB</td>
<td>28 Jul 2021 10:53</td>
</tr>
<tr>
<td>manifest-sha256.txt</td>
<td>7.3 KB</td>
<td>28 Jul 2021 10:53</td>
</tr>
<tr>
<td>manifest-sha512.txt</td>
<td>11 KB</td>
<td>28 Jul 2021 10:53</td>
</tr>
<tr>
<td>metadata.json</td>
<td>9.8 KB</td>
<td>28 Jul 2021 10:53</td>
</tr>
<tr>
<td>tagmanifest-md5.txt</td>
<td>315 B</td>
<td>28 Jul 2021 10:53</td>
</tr>
<tr>
<td>tagmanifest-sha1.txt</td>
<td>363 B</td>
<td>28 Jul 2021 10:53</td>
</tr>
<tr>
<td>tagmanifest-sha256.txt</td>
<td>507 B</td>
<td>28 Jul 2021 10:53</td>
</tr>
<tr>
<td>tagmanifest-sha512.txt</td>
<td>891 B</td>
<td>28 Jul 2021 10:53</td>
</tr>
</tbody>
</table>
NEW FEATURES
HIERARCHICAL DATASETS AND ARCHIVES
03
DATASETS HIERARCHY – EMBEDDED ARCHIVES

Builds hierarchy of datasets which can be used together or independently.
NEW FEATURES
HARD LINKS AND SYMBOLIC LINKS
04
HARD LINKS

Linked files
SYMBOLIC LINKS

Symbolic link to a directory or file
R&D ACHIEVEMENTS

PERFORMANCE IMPROVEMENTS

05
Push Ingestion Speed
# Benefits of Our Approach

## Data Mgmt. & Discovery
- Built around **hybrid** and distributed architecture
- Support well known concepts: **files, folders, links, symbolic links, datasets, archives, metadata**
- Harvesting and complex **indexing** for data **discovery**
- Deep integration with IdPs and multi-level **access control**
- **PUSH** and **PULL** ingestion
- **Open Data** ready

## Data Processing
- Virtual File System **POSIX API**
- Automation Workflows
- Flexible **automation functions** driven by community and end-users
- Out of the box **auto-scalability**
- High **throughput** performance
- No need for expensive temporary **POSIX block storage**

## Licensing & Platform
- **Fully Open Source - MIT**
- Platform **Agnostic**
- **OpenFaaS** support
- Kubernetes ready
- **Backend Storage Agnostic**
**Selected Future Plans**

- Storage grants distributing capacity from procurer to end-user
- UX Improvements e.g. External storage at the user level
- Enhancing performance to the 100 Gbps level
- CTS certification
- Multi-tiered storage management including tape storage
- OAIS Archival Wizard enabling lower entry barrier
- Data Lifecycle Policies
QUESTIONS?