



Onedata FaaS Workflow Engine for Archivisation and beyond

Presented by: Lukasz Dutka

ONEDATA

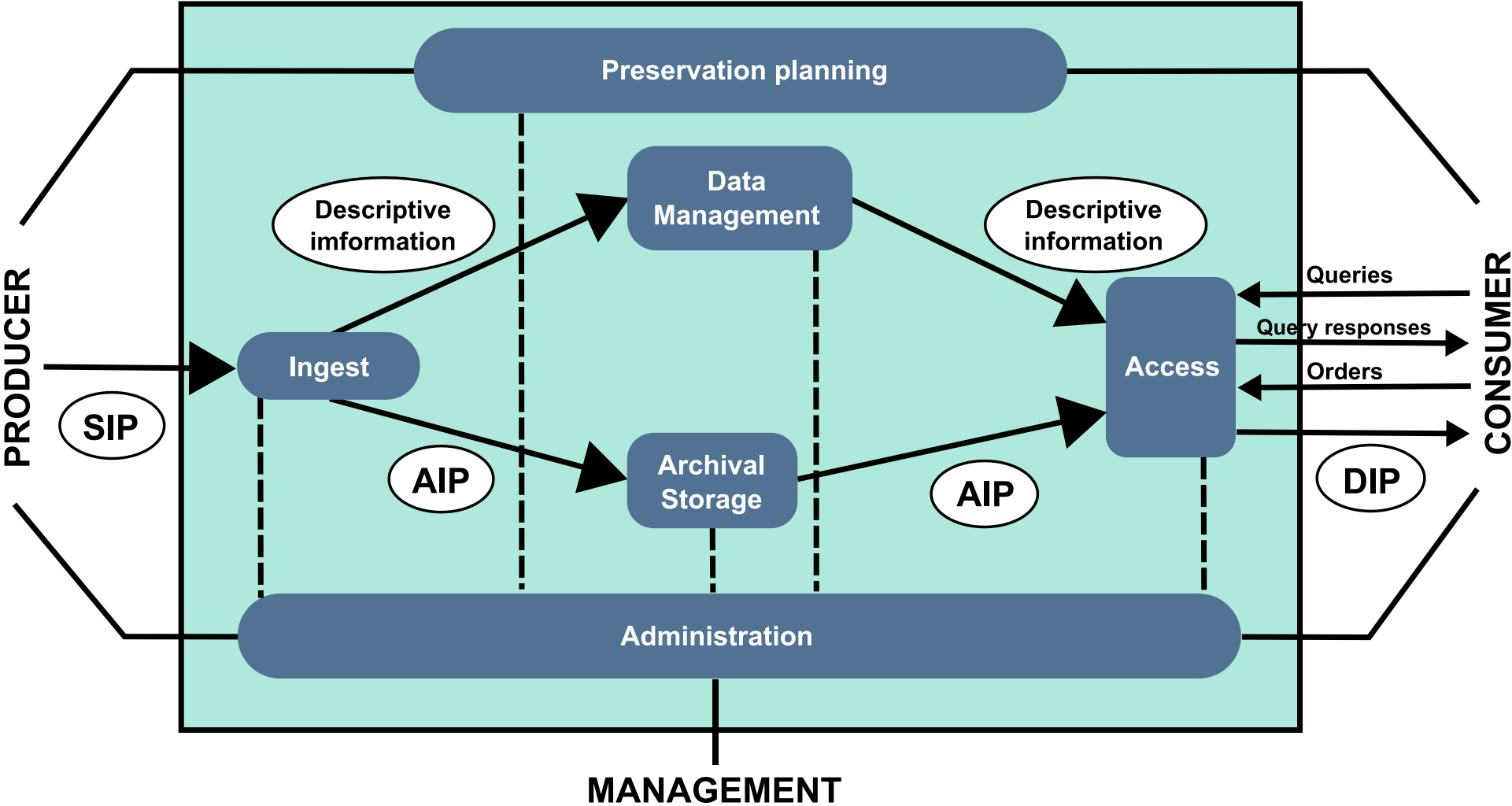


IN COLLABORATION WITH

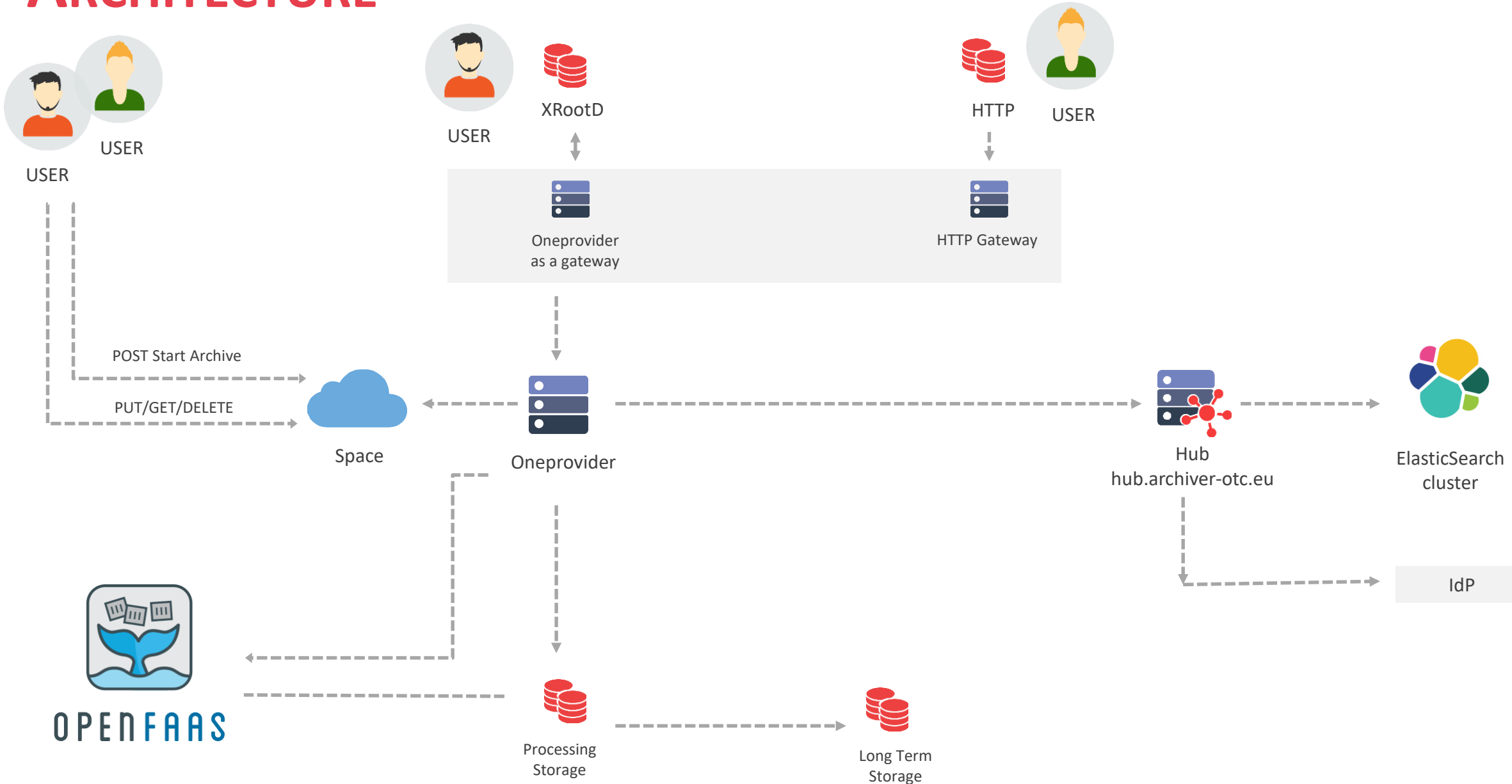
T · · · Systems ·



OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)



ARCHITECTURE



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.

```
$ cat fetch.txt
http://packages.devel.onedata.org/apt/ubuntu/1
02/pool/main/e/erlang/erlang-asn1_20.2.2+dfsg-
ubuntu1ppa6~xenial_amd64.deb 756358 data/dir1/
dir12/erlang-asn1_20.2.2+dfsg-0ubuntu1ppa6~xeni
l_amd64.deb
http://packages.devel.onedata.org/apt/ubuntu/1
02/pool/main/e/erlang/erlang-base-hipe_20.2.2+
dfsg-0ubuntu1ppa6~xenial_amd64.deb 9136096 data
erlang-base-hipe_20.2.2+dfsg-0ubuntu1ppa6~xeni
l_amd64.deb
http://packages.devel.onedata.org/apt/ubuntu/1
02/pool/main/e/erlang/erlang-base_20.2.2+dfsg-
ubuntu1ppa6~xenial_amd64.deb 7386200 data/dir1
dir12/erlang-base_20.2.2+dfsg-0ubuntu1ppa6~xer
al_amd64.deb
http://packages.devel.onedata.org/apt/ubuntu/1
02/pool/main/e/erlang/erlang-common-test_20.2.
+dfsg-0ubuntu1ppa6~xenial_amd64.deb 1065320 da
a/dir1/erlang-common-test_20.2.2+dfsg-0ubuntu1
ppa6~xenial_amd64.deb
```



NEW FEATURES

AUTOMATION ENGINE

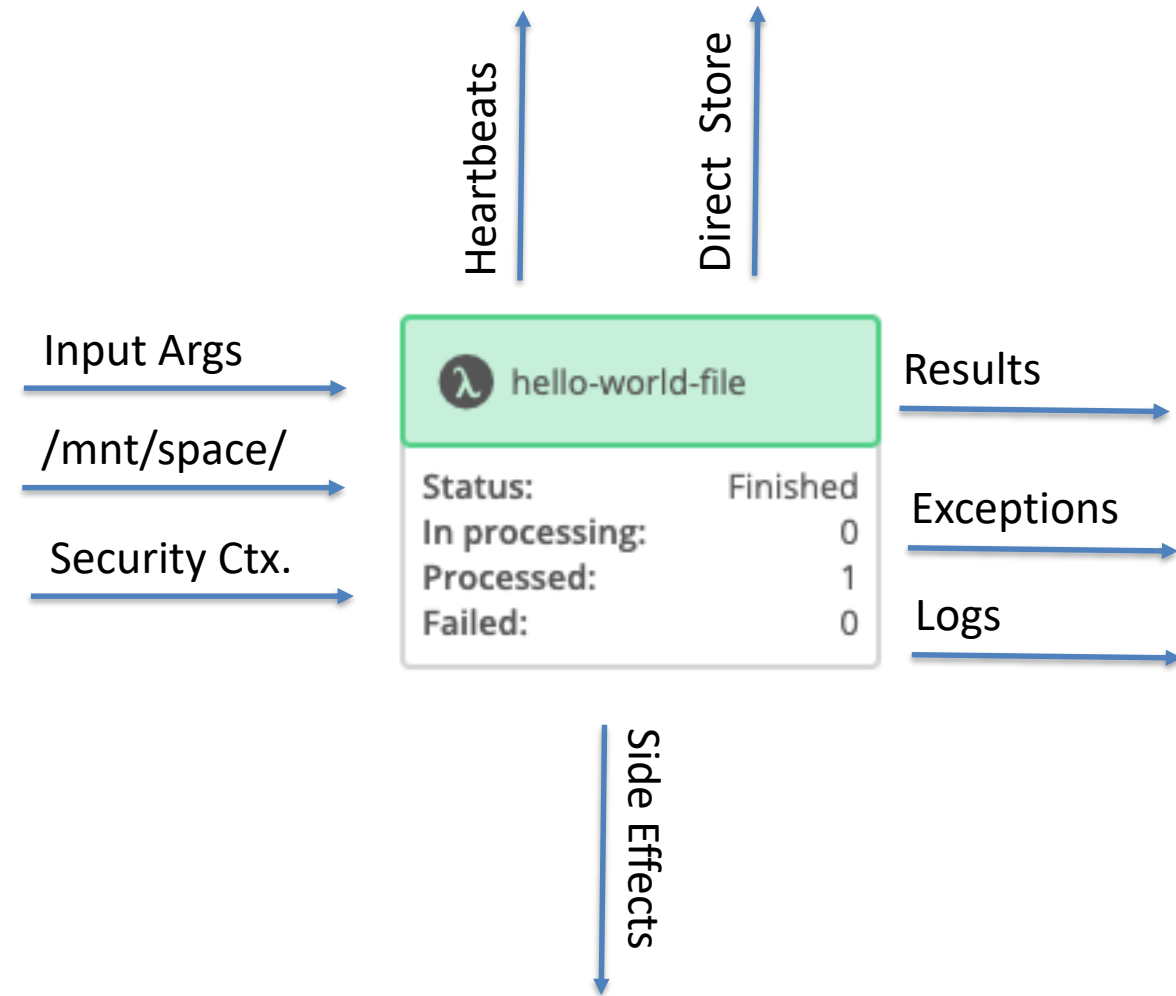
01

FAAS LAMBDA

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

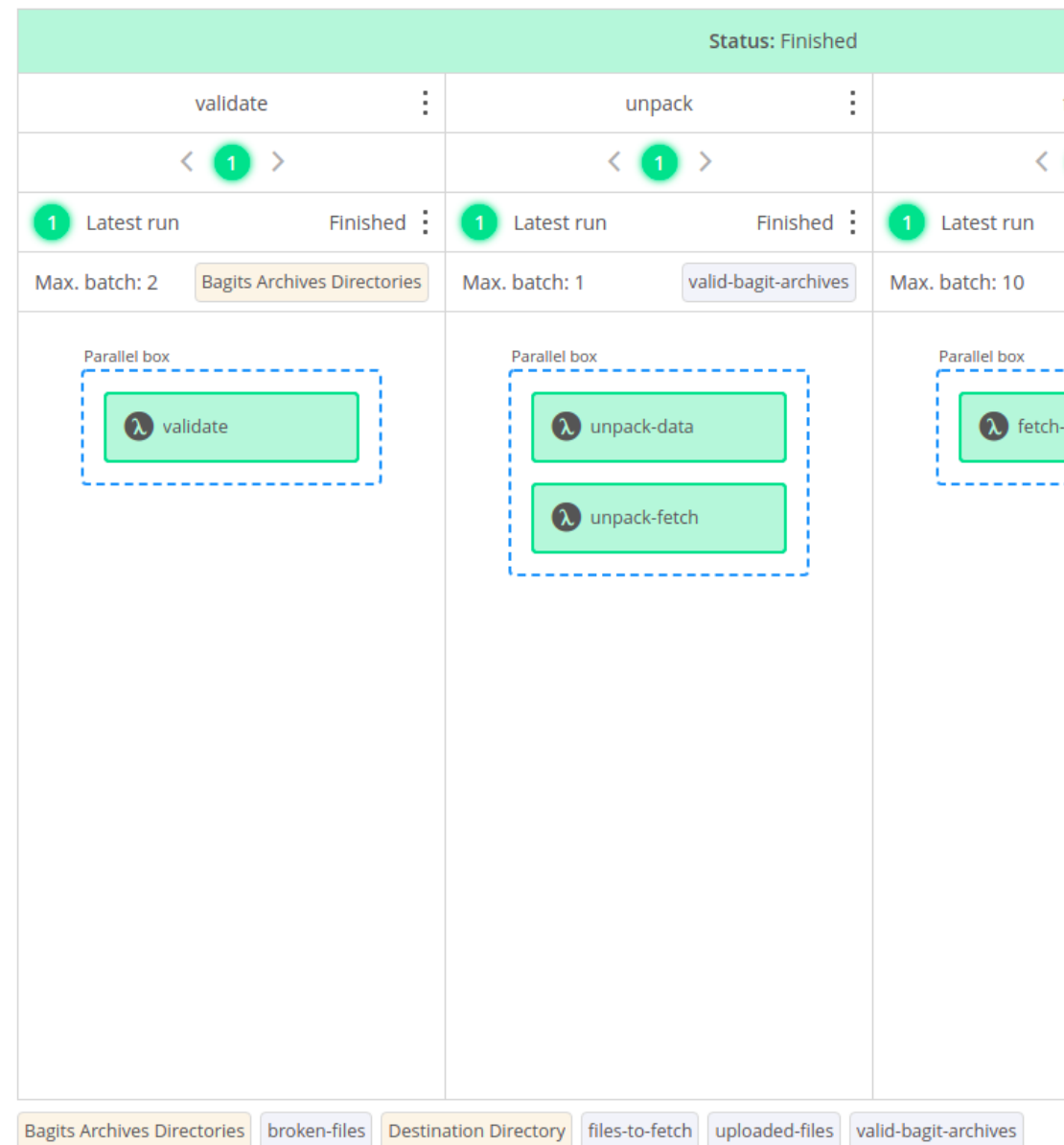


WORKFLOW

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

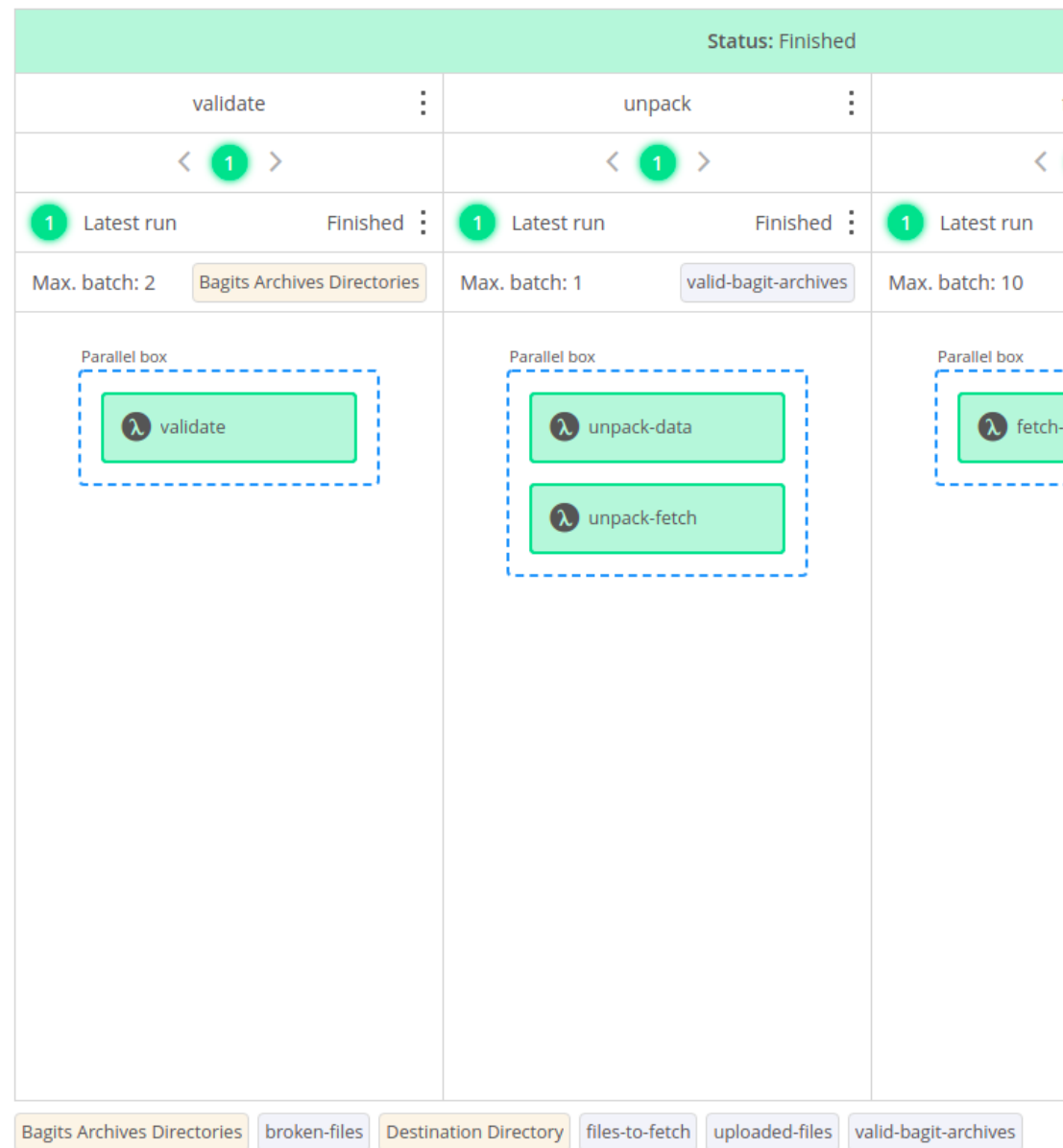
- Can be exported to JSON and reused by someone else



STORE

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

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

The screenshot displays the 'AUTOMATION' section of the Inventory application. A vertical sidebar on the left contains navigation icons, with the 'Automation' icon highlighted in red. The main content area shows a list of automation workflows:

- Lukasz Inventory
- System Inventory** (highlighted in red)
- Workflows
- Lambdas
- Members

Below the list, the 'WORKFLOWS' section is visible, showing a search bar and three workflow cards:

- Bagit Extractor**: Automation workflow processing. Table with 1 revision (Stable).
- Check Format**: Automation workflow for checking. Table with 2 revisions (Draft, Stable).
- Checksum Calculator via POSIX**: Calculate checksums of file using. Table with 1 revision (Stable).
- Hello World File**: The simplest possible function w. Table with 1 revision (Stable).

Rev.	State	Description
1	Stable	Stable Bagit Extra

Rev.	State	Description
2	Draft	Detecting new file
1	Stable	First version

Rev.	State	Description
1	Stable	Added MD5, SHA2

Rev.	State	Description
1	Stable	First version



DEMO



NEW FEATURES

OAIS ARCHIVISATION

02

DATASETS, AIP, DIP, INCREMENTAL ARCHIVES

The screenshot displays the OneDrive web interface. On the left is a sidebar menu with options: Overview, Files, Shares, Transfers, **Datasets**, Providers, Members, Harvesters, and Automation. The main area shows a collection named 'My Collection' with tabs for 'AIP' and 'DIP'. Below the tabs, a view provided by 'OTC Test 1' is shown. A file collection is displayed with a table of files:

Files	Size	Modification
bagit.txt	53 B	28 Jul 2021 10:53
data	—	28 Jul 2021 10:53
manifest-md5.txt	5.8 KiB	28 Jul 2021 10:53
manifest-sha1.txt	6.2 KiB	28 Jul 2021 10:53
manifest-sha256.txt	7.5 KiB	28 Jul 2021 10:53
manifest-sha512.txt	11 KiB	28 Jul 2021 10:53
metadata.json	9.8 KiB	28 Jul 2021 10:53
tagmanifest-md5.txt	315 B	28 Jul 2021 10:53
tagmanifest-sha1.txt	363 B	28 Jul 2021 10:53
tagmanifest-sha256.txt	507 B	28 Jul 2021 10:53
tagmanifest-sha512.txt	891 B	28 Jul 2021 10:53

Red annotations include an arrow pointing to the 'AIP' and 'DIP' tabs labeled 'AIP or DIP selector', and another arrow pointing to the context menu of a file in the table.



NEW FEATURES

HIERARCHICAL DATASETS AND ARCHIVES

03

DATASETS HIERARCHY – EMBEDDED ARCHIVES

DATA

Search...

- 1000 Genoms (810 TiB, 3)
- dataset-lookup (500 TiB, 1)
- Demo Books Collection (1000 TiB, 2)
- Demo Hierarchy** (1000 TiB, 2)

Overview
Files
Shares
Transfers
Datasets
Providers
Members
Harvesters
Automation

DATASETS Attached Detached

View provided by OTC Test 1. Choose other Oneprovider...

CREATE ARCHIVE

CC

Description: Archive of Hierarchical Dataset ✓

Layout: plain bagit

Create nested archives: ← Builds hierarchy of datasets which can be used together of independently

Incremental:

Include DIP:

Cancel Create



NEW FEATURES

HARD LINKS AND SYMBOLIC LINKS

04

HARD LINKS

The screenshot displays a file management interface with a sidebar on the left and a main content area on the right. The sidebar includes navigation options: Overview, Files, Shares, Transfers, Datasets, Providers, Members, Harvesters, and Automation. The main content area shows a list of files and folders. A red arrow points from the text "Linked files" to the "6 hard links" button for the file "op-worker_21.02.0.alpha13.207.gcff6689.orig.tar.gz".

Item	Hard Links	Meta	Size	Date
dir1	—	—	—	28 Jul 2021 10:53
dir2	—	—	—	28 Jul 2021 10:53
libradosstriper1_14.2.2-1xenial_amd64.deb	6 hard links	Meta	325.9 KiB	28 Jul 2021 10:51
librbd1_14.2.2-1bionic_amd64.deb	6 hard links	Meta	1.4 MiB	28 Jul 2021 10:51
librgw2_14.2.2-1bionic_amd64.deb	6 hard links	Meta	3.5 MiB	28 Jul 2021 10:51
metadata.json	6 hard links	Meta	6.2 KiB	28 Jul 2021 10:51
op-worker_21.02.0.alpha13.207.gcff6689.orig.tar.gz	6 hard links	Meta	172.7 MiB	28 Jul 2021 10:51
op-worker_21.02.0.alpha13.242.g50a0df5-1~bionic_amd64.deb	6 hard links	Meta	386 B	28 Jul 2021 10:51
oz-panel_21.02.0.alpha13.38.g7aa9097-1~bionic.diff.gz	6 hard links	Meta	381 B	28 Jul 2021 10:51
oz-worker_21.02.0.alpha13.28.g4f6d523-1~bionic.diff.gz	6 hard links	Meta	383 B	28 Jul 2021 10:51
oz-worker_21.02.0.alpha13.28.g4f6d523-1~bionic.deb	6 hard links	Meta	370 B	28 Jul 2021 10:51

SYMBOLIC LINKS

The screenshot shows a file management interface with a sidebar on the left and a main content area on the right. The sidebar contains a search bar and a list of collections: 1000 Genoms (810 TiB), dataset-lookup (500 TiB), Demo Books Collection (1000 TiB), and Demo R (1000 TiB). Below the collections are navigation options: Overview, Files, Shares, Transfers, Datasets, Providers, Members, Harvesters, and Automation.

The main content area shows a view provided by OTC Test 1. The breadcrumb path is: 28 Jul 2021 11:03 — Hierarchical Archive / data / My Collection / dir1. Below the breadcrumb is a table of files and directories.

Files	Size	Modification
aws-c-event-stream_0.1.5-0ubuntu1ppa1-xenial_amd64.deb	Meta 19 KIB	28 Jul 2021 11:03
curl_7.58.0-1ubuntu3-xenial.debian.tar.xz	Meta 33.4 KiB	28 Jul 2021 11:03
curl_7.58.0-1ubuntu3-xenial.dsc	Meta 2.4 KIB	28 Jul 2021 11:03
dir11	—	28 Jul 2021 11:03
dir12	—	28 Jul 2021 11:03
libcurl4-gnutls-dev_7.58.0-1ubuntu3-xenial_amd64.deb	Meta 379.3 KIB	28 Jul 2021 11:03
libcurl4-nss-dev_7.58.0-1ubuntu3-xenial_amd64.deb	Meta 385.2 KIB	28 Jul 2021 11:03
libcurl4-openssl-dev_7.58.0-1ubuntu3-xenial_amd64.deb	Meta 379.9 KIB	28 Jul 2021 11:03
libcurl4_7.58.0-1ubuntu3-xenial_amd64.deb	Meta 296.5 KIB	28 Jul 2021 11:03
librados2_14.2.2-1bionic_amd64.deb	Meta 2.9 MiB	28 Jul 2021 11:03
oz-worker_21.02.0.alpha13.30.g0aabdf9-1-bionic_amd64.deb	Meta 385 B	28 Jul 2021 11:03
oz-worker_21.02.0.alpha13.52.g5bfcaca.orig.tar.gz	Meta 378 B	28 Jul 2021 11:03

A red arrow points from the text "Symbolic link to a directory or file" to the entry "dir11" in the table.

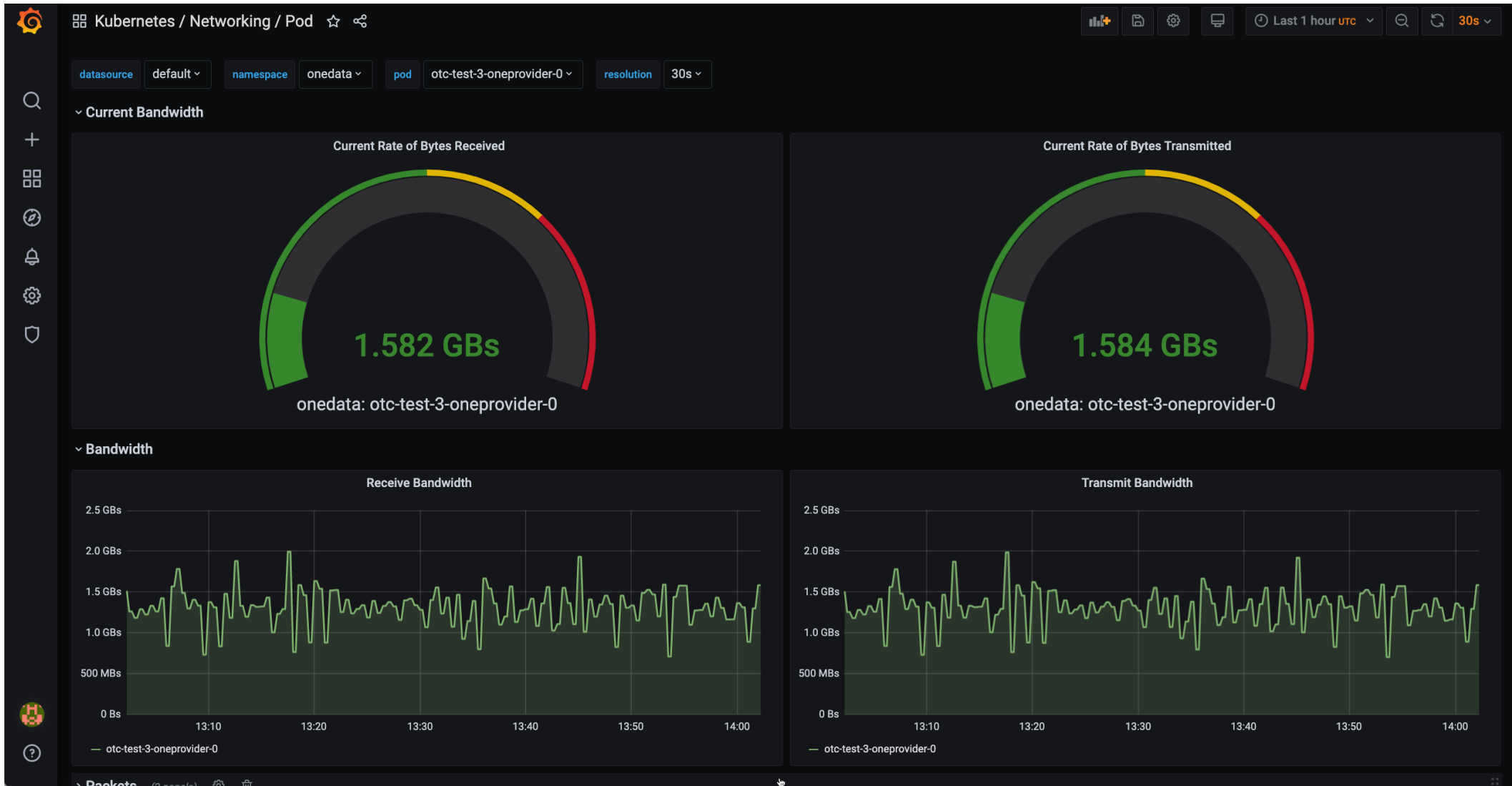


R&D ACHIEVEMENTS

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

ONE DATA

QUESTIONS?