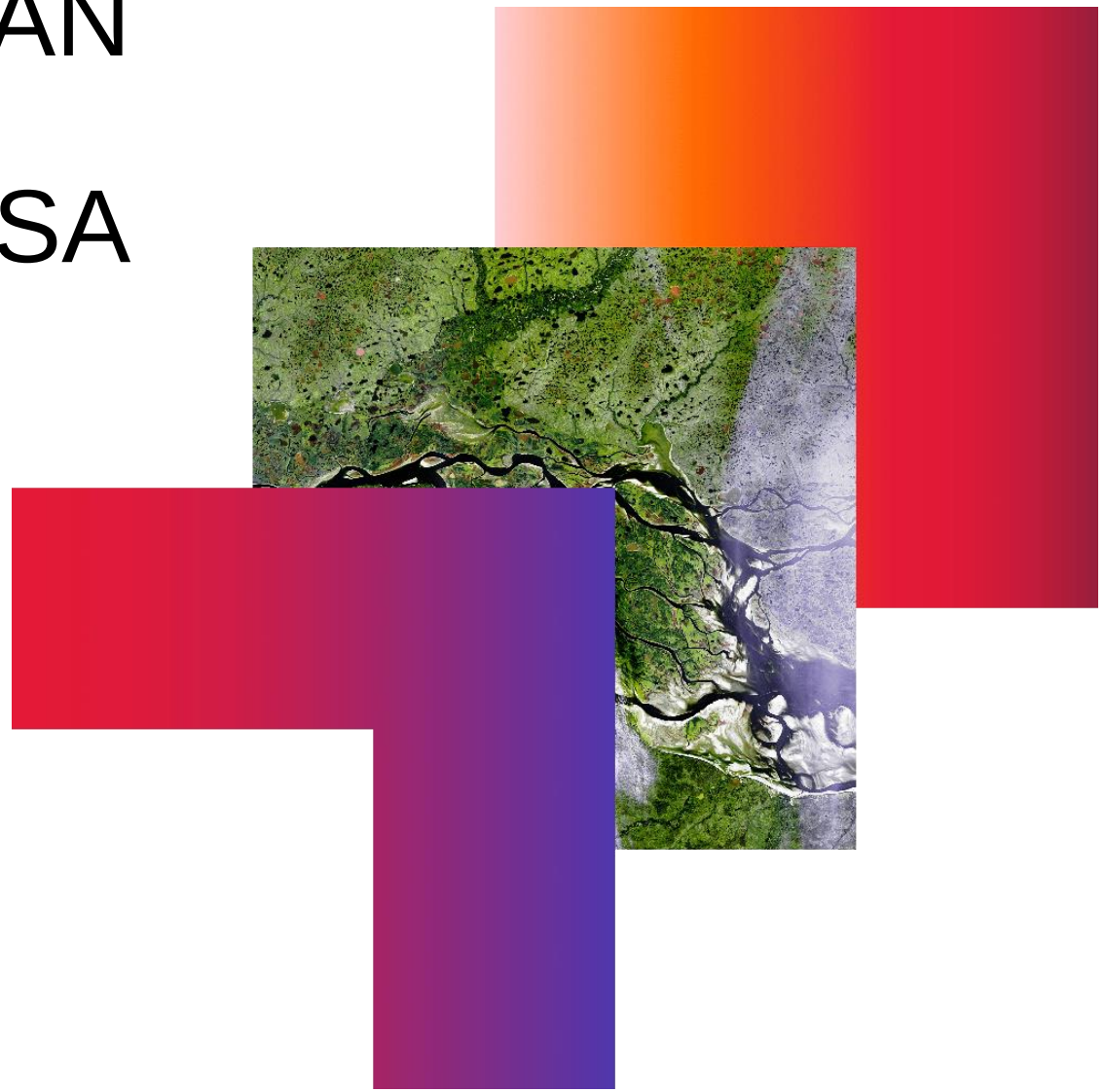
ESA Datalabs – PIPEMAN

Pipeline computing in ESA Datalabs

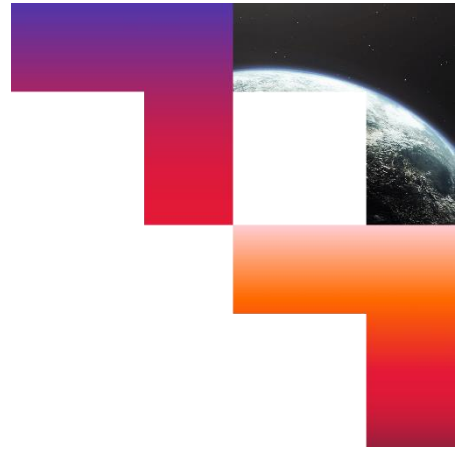


Brief overview – 24.11.2022

Alo Joosepson

CGI Estonia

Agenda



1. What is a pipeline?
2. User profiles
3. Logical architecture
4. Main functions
5. Release plans
6. Hands-on session invitation

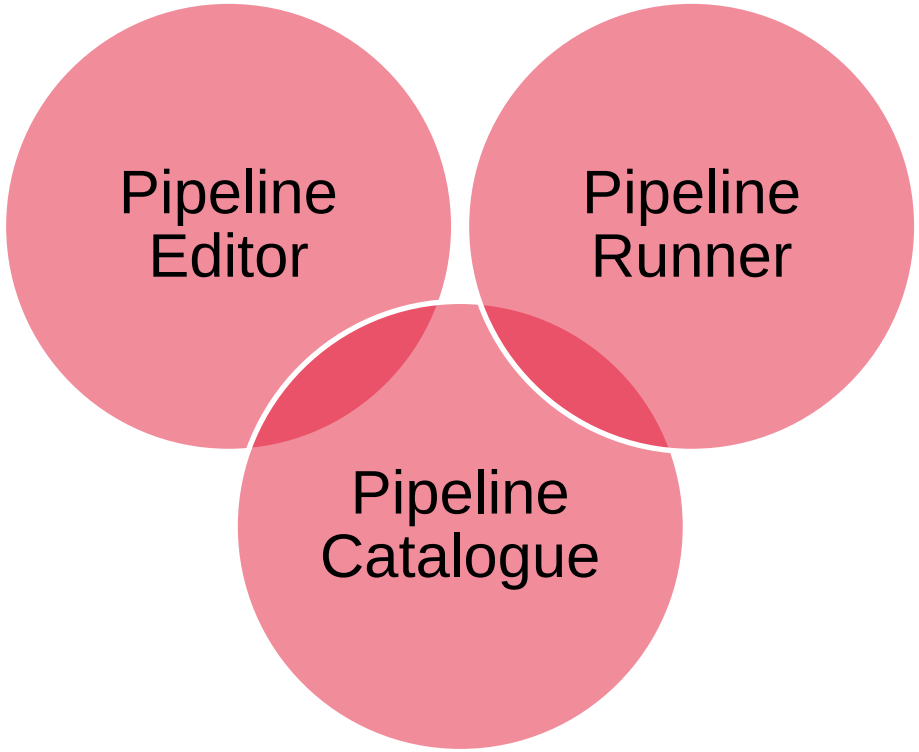
What is a pipeline?

A predefined data processing workflow

- Takes one or more inputs (files, parameters, messages, data from databases)
- Processes them in a series of one or more steps.
- Produces one or more outputs (files, messages, data into databases).
- Repeatable
- Re-usable
- Versionable
- Can do the heavy lifting of massive processing of typical data

User profiles

Logical architecture



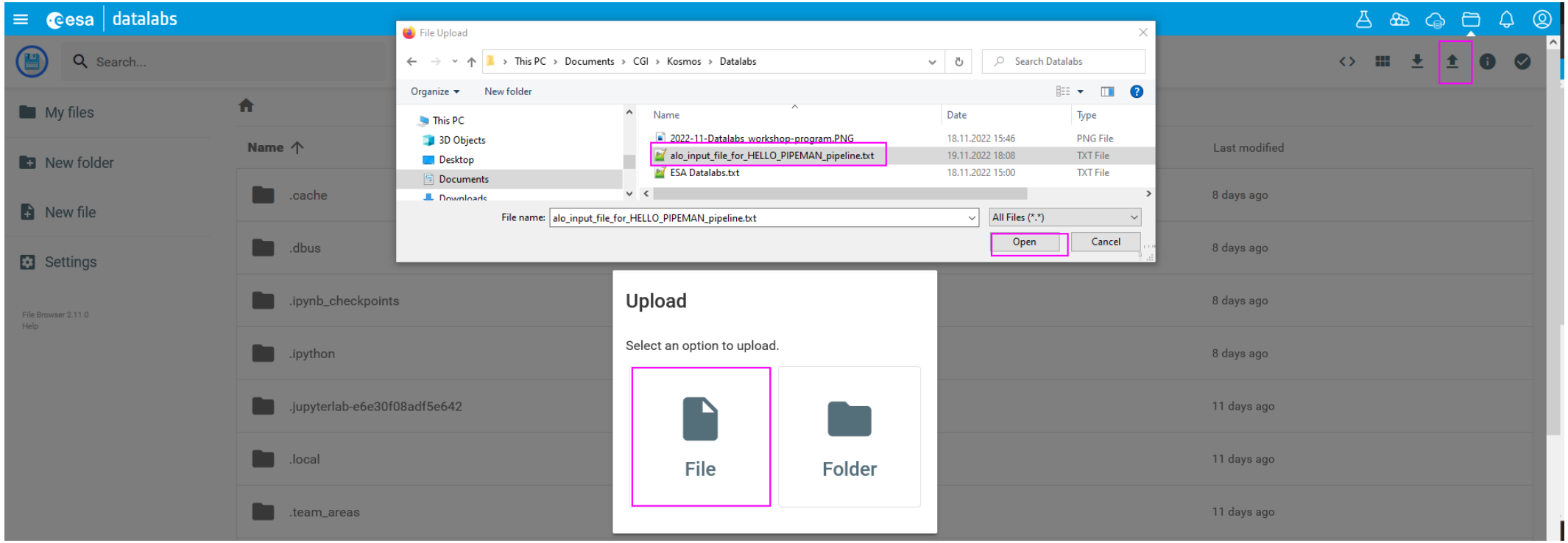
Authentication and user rights

Data Discovery

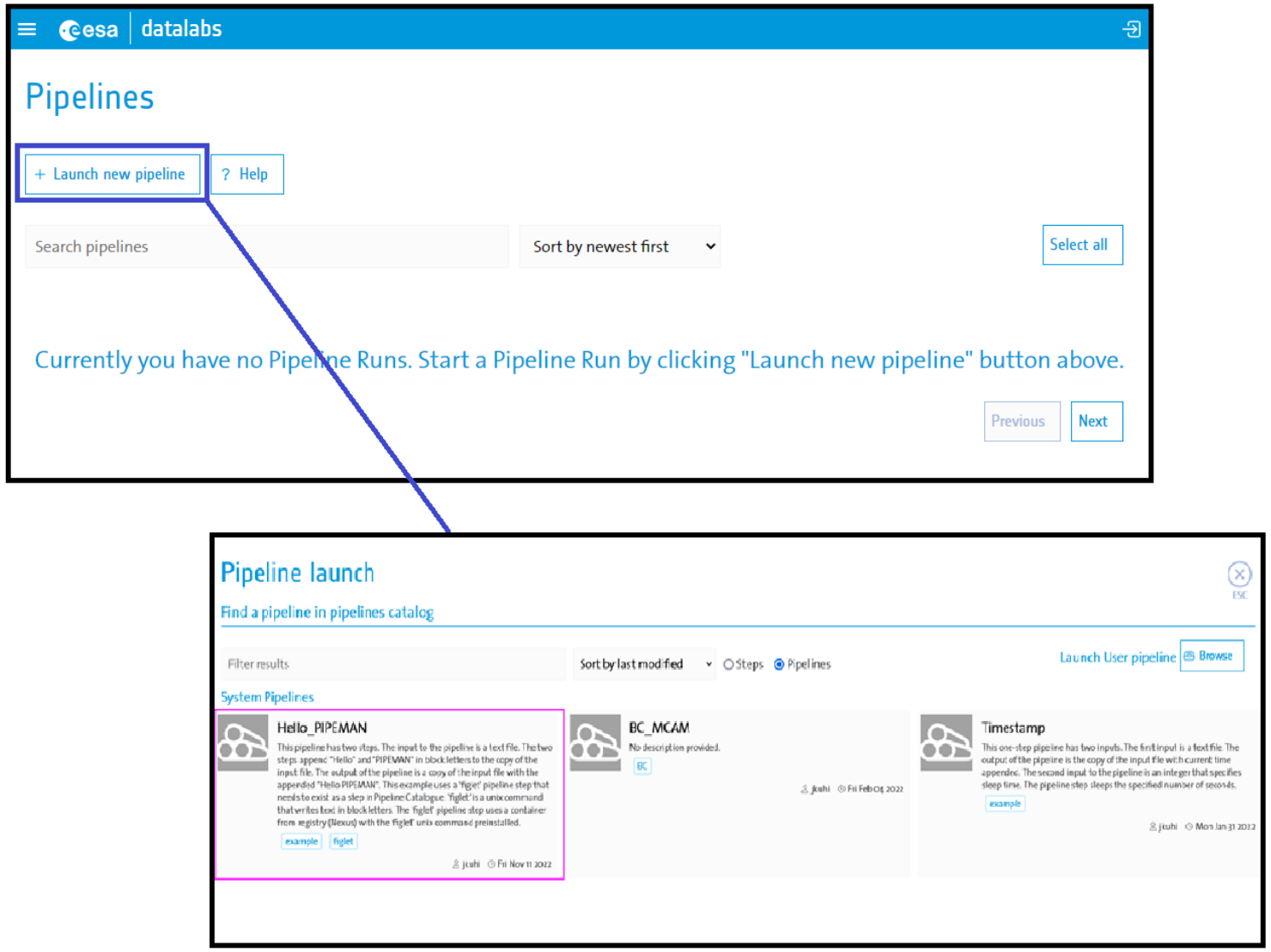
Data Storage

Datalabs

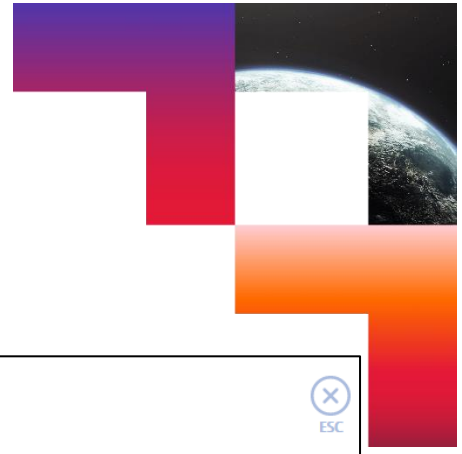
Pipeline Runner – Create, find or upload input data

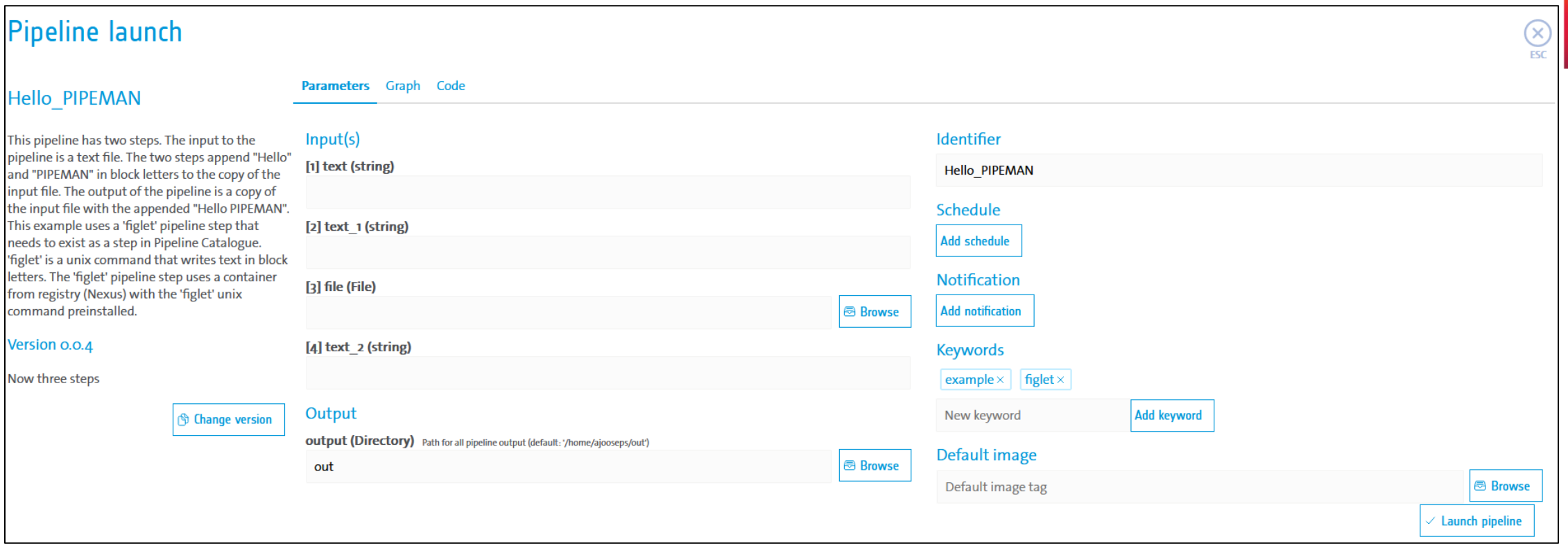


Pipeline Runner – Find a pipeline

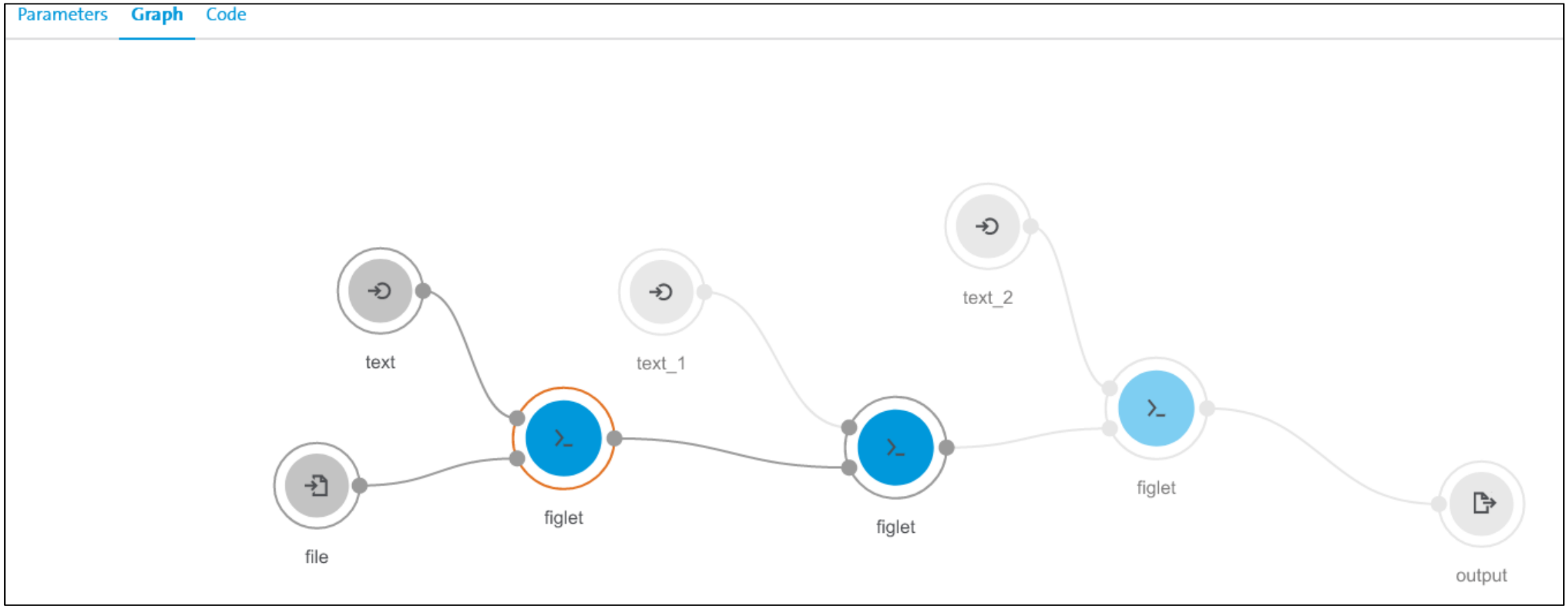


Pipeline Runner – Launch parameters

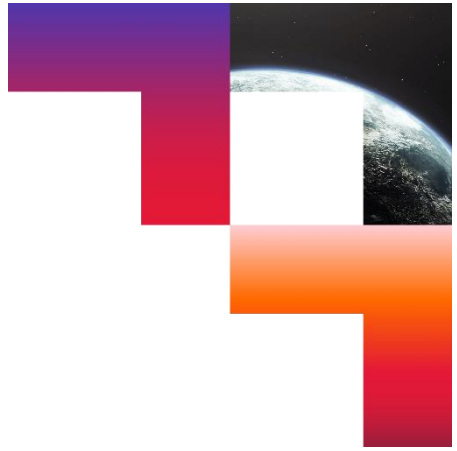




Pipeline Runner – Graph



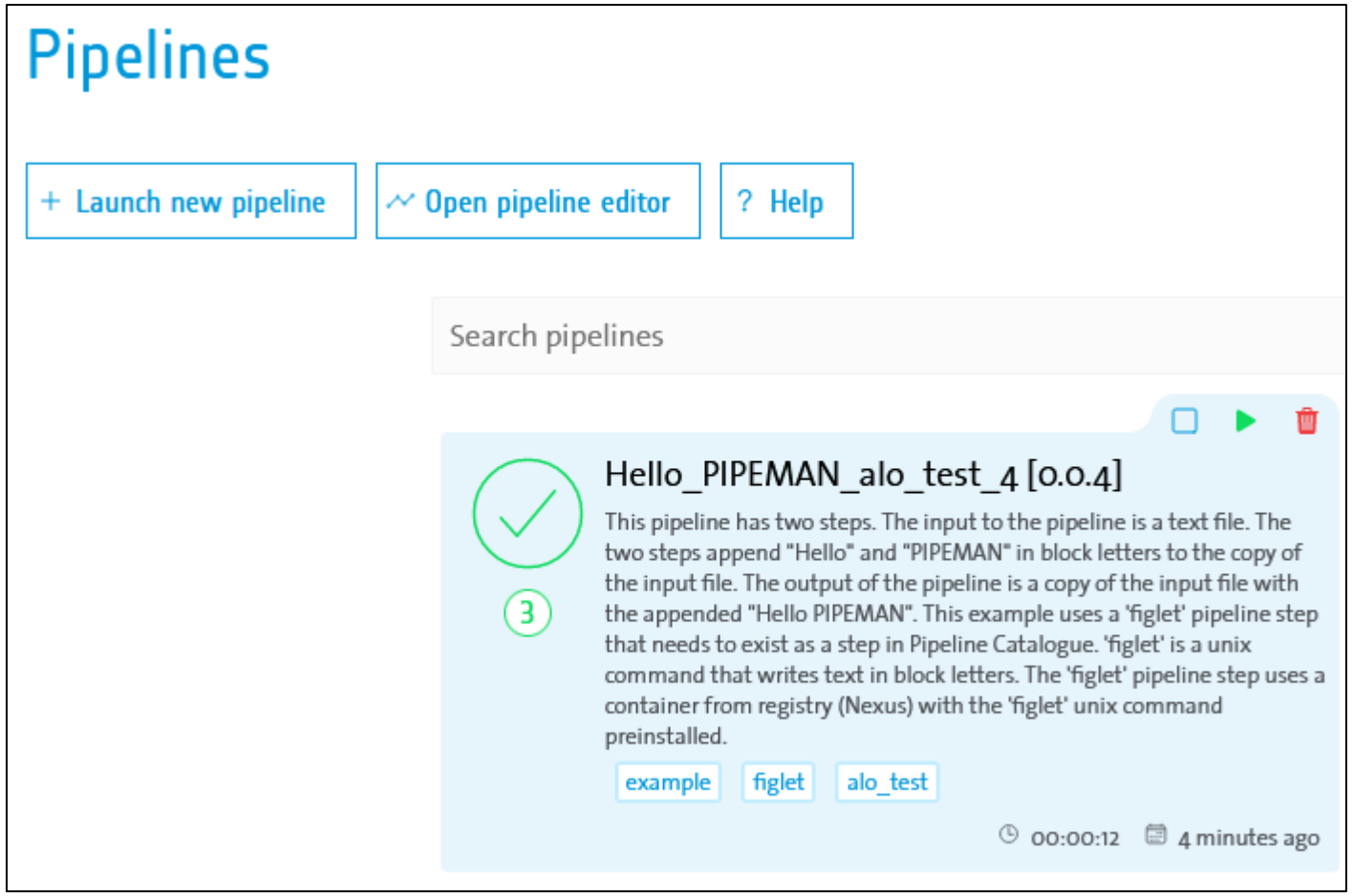
Pipeline Runner – Pipeline Run Code



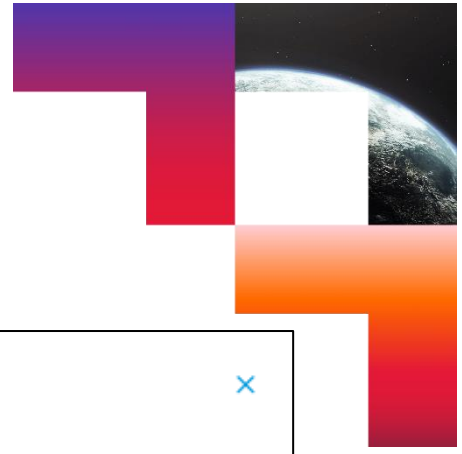
Parameters Graph **Code**

```
class: Workflow
cwlVersion: v1.0
id: _hello__p_i_p_e_m_a_n
label: Hello_PIPEMAN
$namespaces:
  sepp: 'https://localhost/'
  sbg: 'https://www.sevenbridges.com/'
inputs:
  - id: text
    type: string
    'sbg:x': -1083.9599609375
    'sbg:y': -397.1409912109375
  - id: text_1
    type: string
    'sbg:x': -839
    'sbg:y': -396
  - id: data
    type: File
    label: file
    'sbg:x': -1138
    'sbg:y': -228.1409912109375
  - id: text_2
    type: string
    'sbg:x': -556
    'sbg:y': -453
outputs:
  - id: output
```

Pipeline Runner – Pipeline Run List



Pipeline Runner – Pipeline Run Execution Statistics



Pipeline run details - Hello_PIPEMAN_alo_test_4 [0.0.4]

Execution | Logs | Parameters | Graph | Code

Status

Progress: Step 3 of 3

Started

11/18/2022, 3:19:10 PM

Finished

11/18/2022, 3:19:23 PM

Schedule

No schedule set

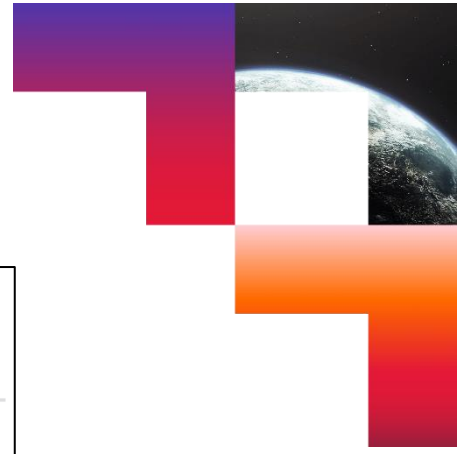
Notifications

Notifying **alo.joosepson@cgi.com** when **pipeline completes**

Keywords

example | figlet | alo_test

Pipeline Runner – Pipeline Run Execution Logs



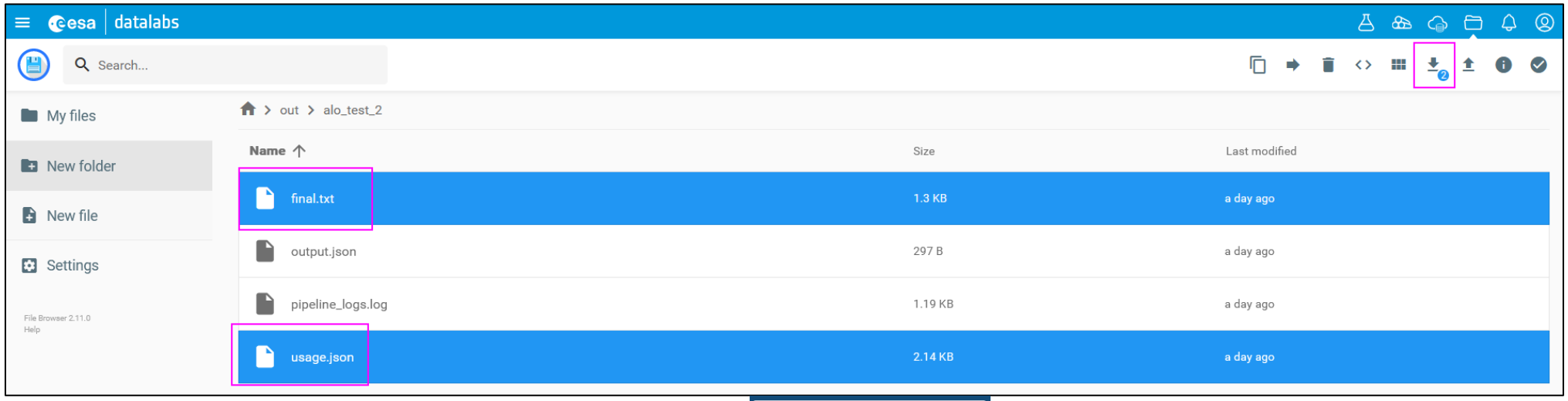
Pipeline run details - Hello_PIPEMAN_alo_test_4 [0.0.4]

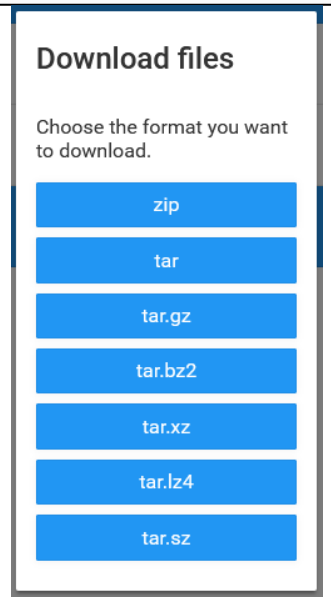
Execution | **Logs** | Parameters | Graph | Code

Pipeline logs

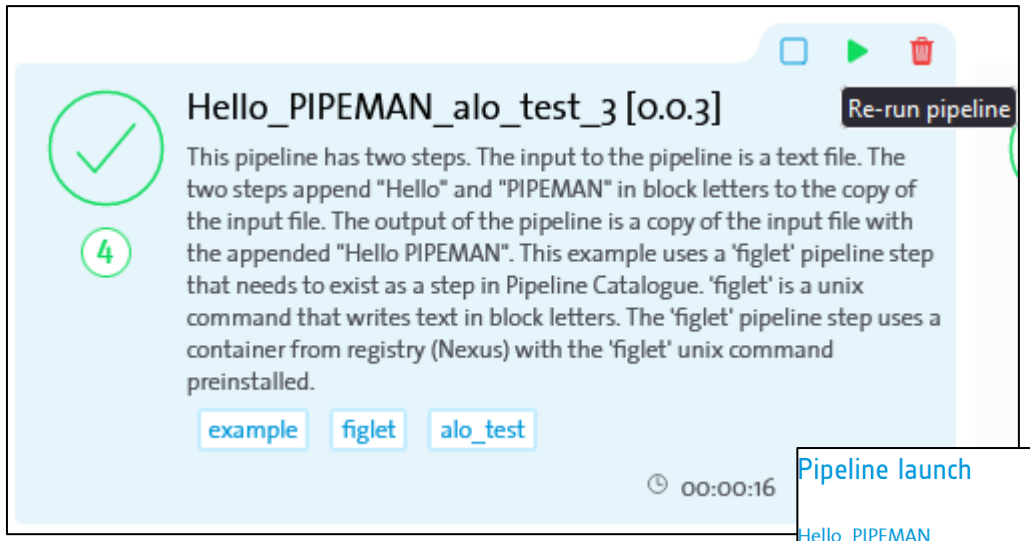
```
1  [2022-11-18 15:19:10] INFO calrissian 0.10.0 (cwltool 3.0.20210124104916)
2  [2022-11-18 15:19:10] INFO Resolved '/cwl/tmp/cwl/Hello_PIPEMAN.pipeline.cwl' to 'file:///cwl/tmp/cwl/H
3  [2022-11-18 15:19:12] INFO [workflow ] starting step figlet
4  [2022-11-18 15:19:12] INFO [step figlet] start
5  [2022-11-18 15:19:12] INFO [workflow ] start
6  [figlet] logs start
7  [figlet] logs end
8  [2022-11-18 15:19:15] INFO [step figlet] completed success
9  [2022-11-18 15:19:15] INFO [workflow ] starting step figlet_1
10 [2022-11-18 15:19:15] INFO [step figlet_1] start
11 [figlet_1] logs start
12 [figlet_1] logs end
13 [2022-11-18 15:19:19] INFO [step figlet_1] completed success
14 [2022-11-18 15:19:19] INFO [workflow ] starting step figlet_2
15 [2022-11-18 15:19:19] INFO [step figlet_2] start
16 [figlet_2] logs start
```

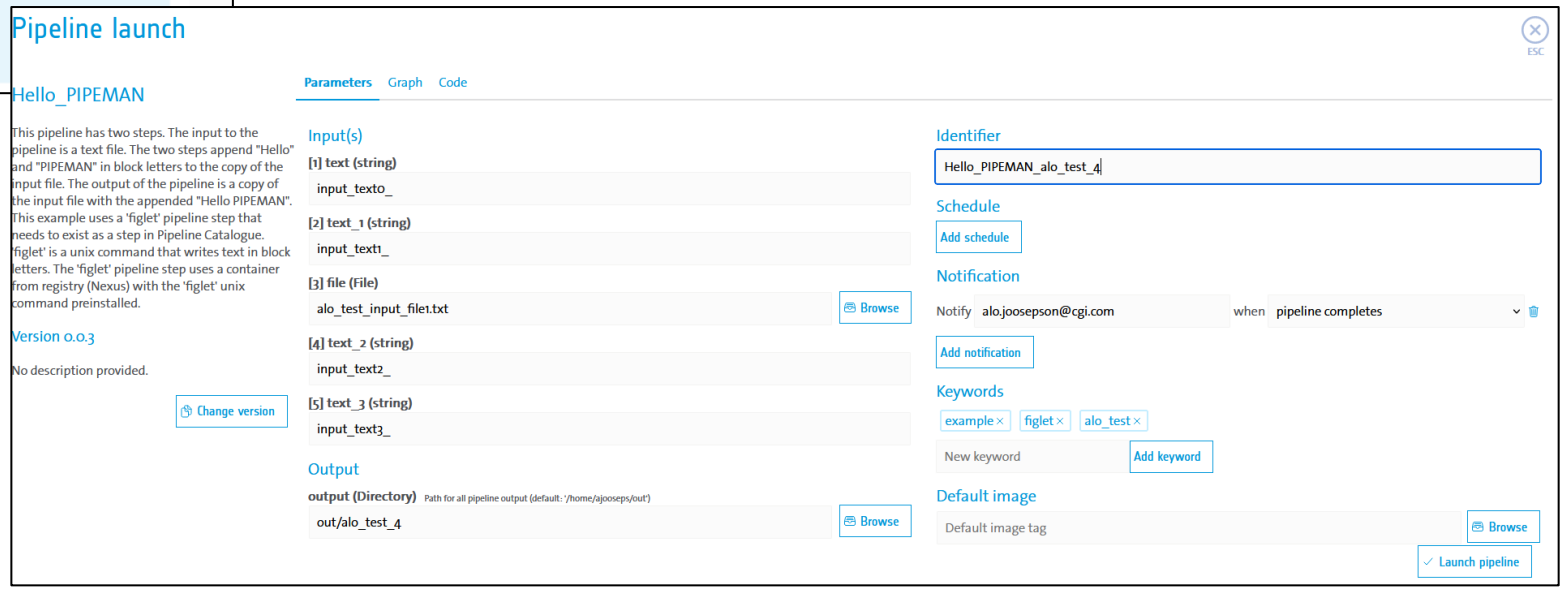
Pipeline Runner – Download results file(s)



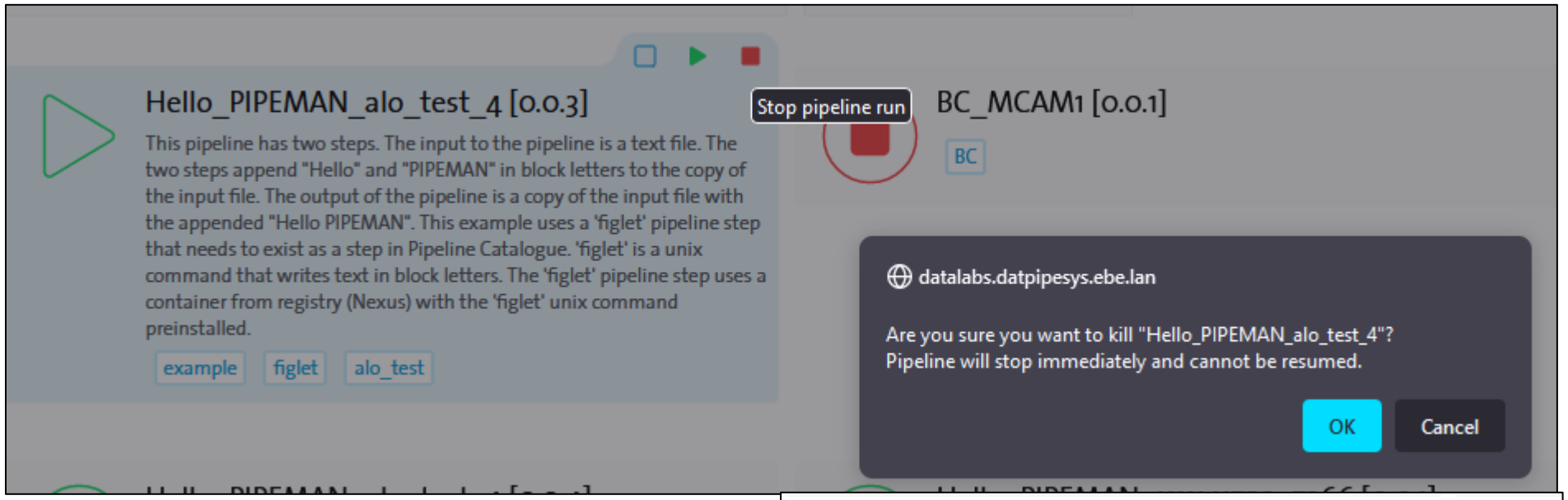


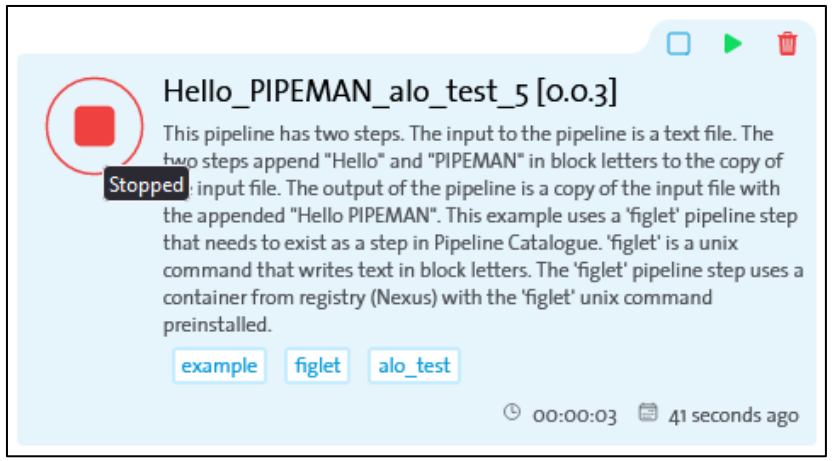
Pipeline Runner – Re-run

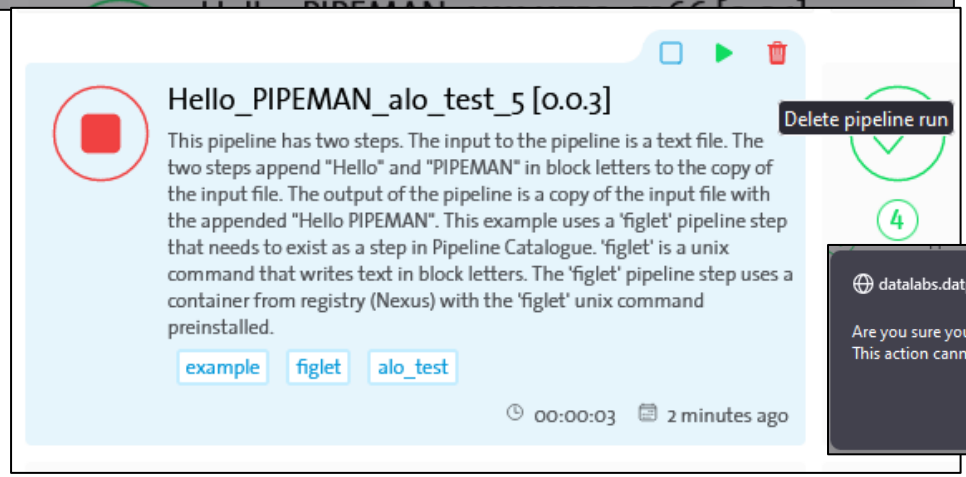




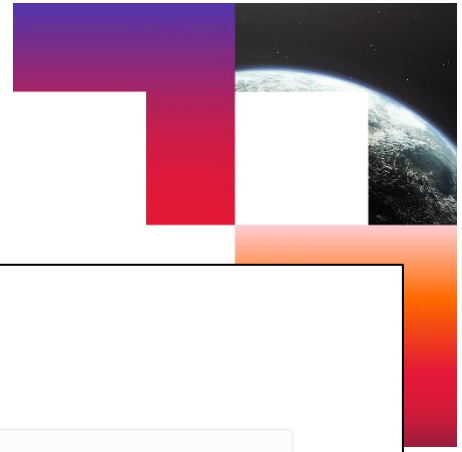
Pipeline Runner – Stop and delete a pipeline run







Pipeline Runner – Scheduled runs and notifications



Schedule

Schedule execution based on following cron expression

25-30 12 25 11 FRI 2022 | Custom

Every minute between 12:25 and 12:30, on day 25 of the month, and on Friday, only in November, only in 2022 (UTC)

Notification

Notify alo.joosepson@cgi.com when Select condition

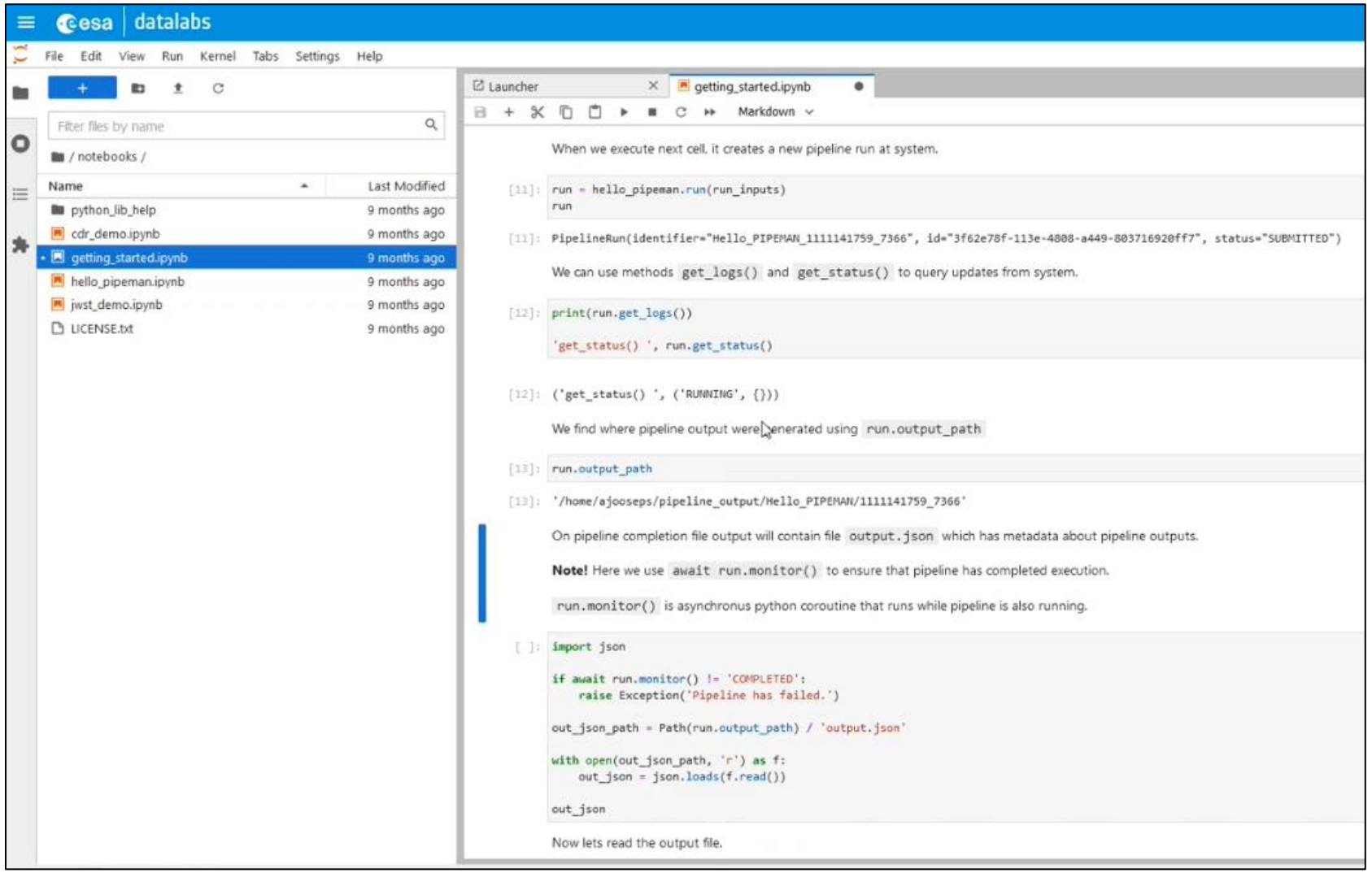
Add notification

Select condition
pipeline fails
pipeline completes
pipeline exceeds time limit of

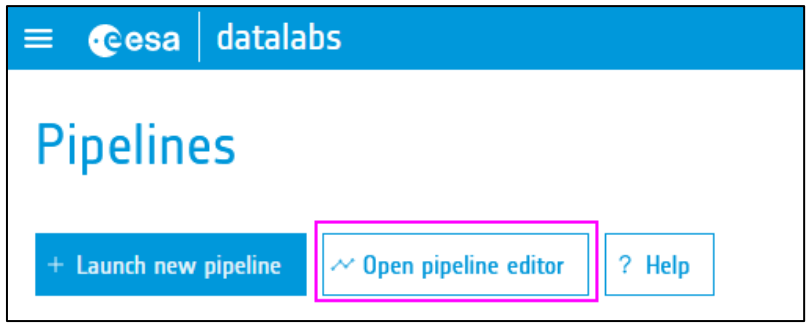
Keywords

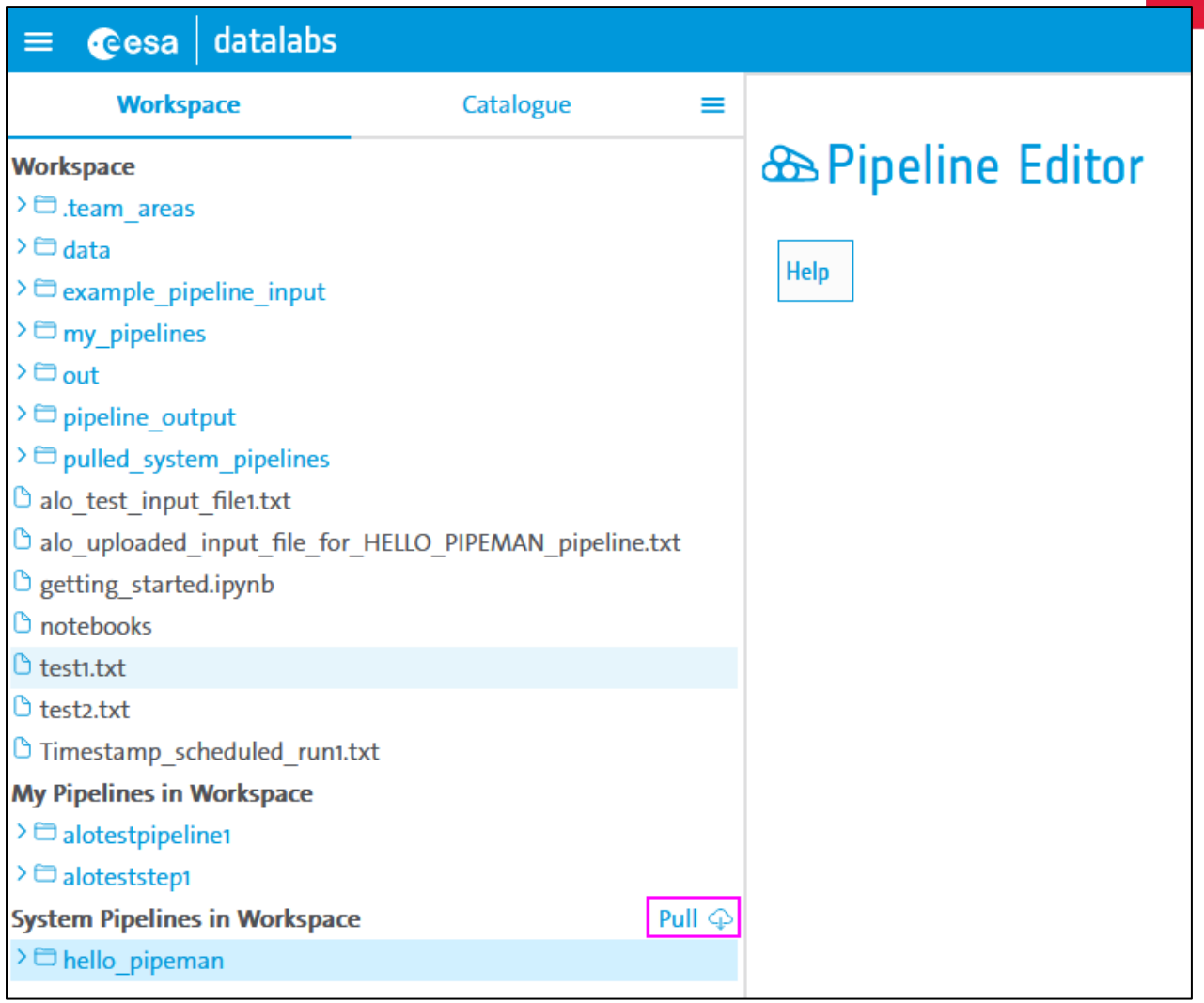
example × figlet ×

Pipeline Runner – Run in a Jupyterlab notebook

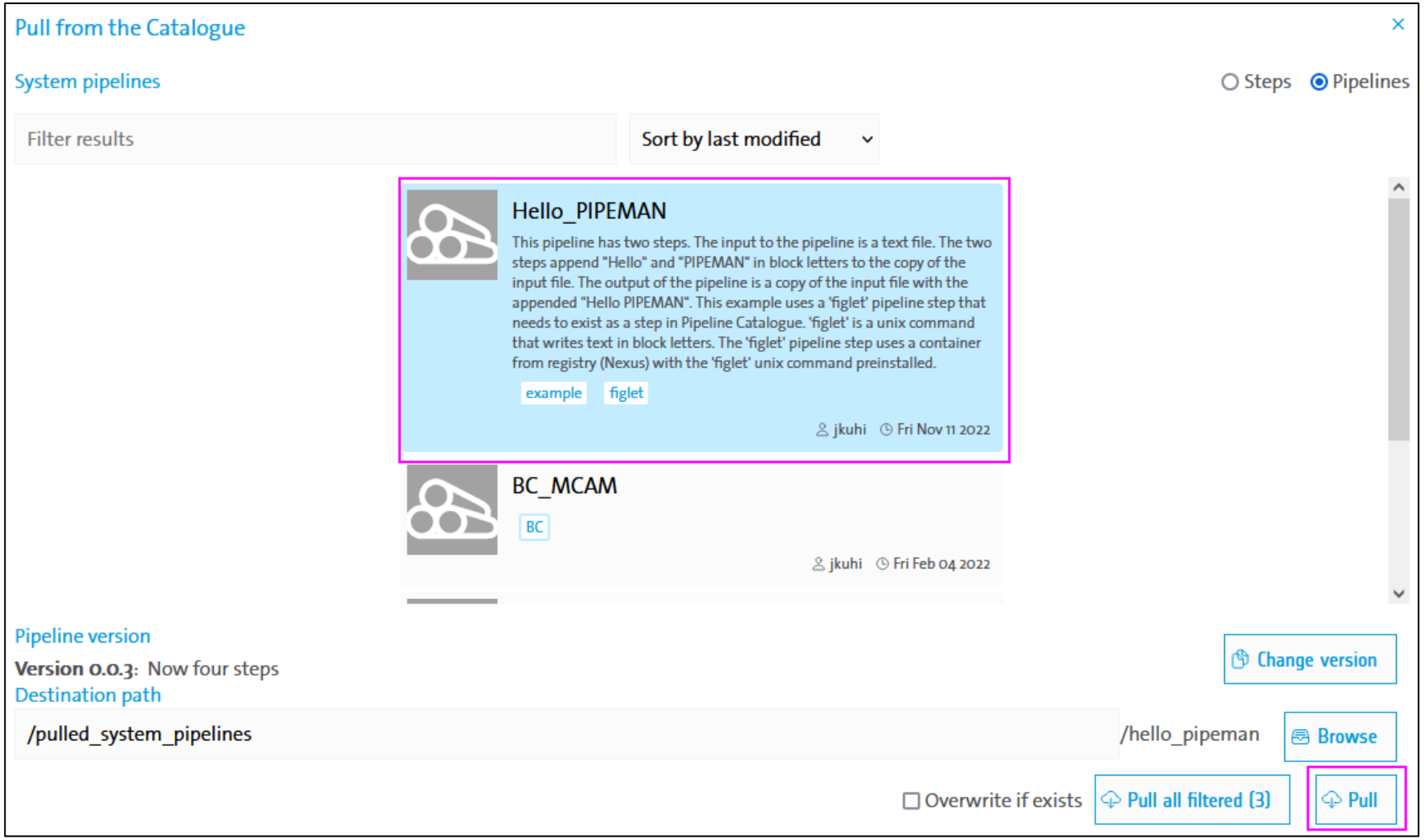


Pipeline Editor – Open, Personal Workspace

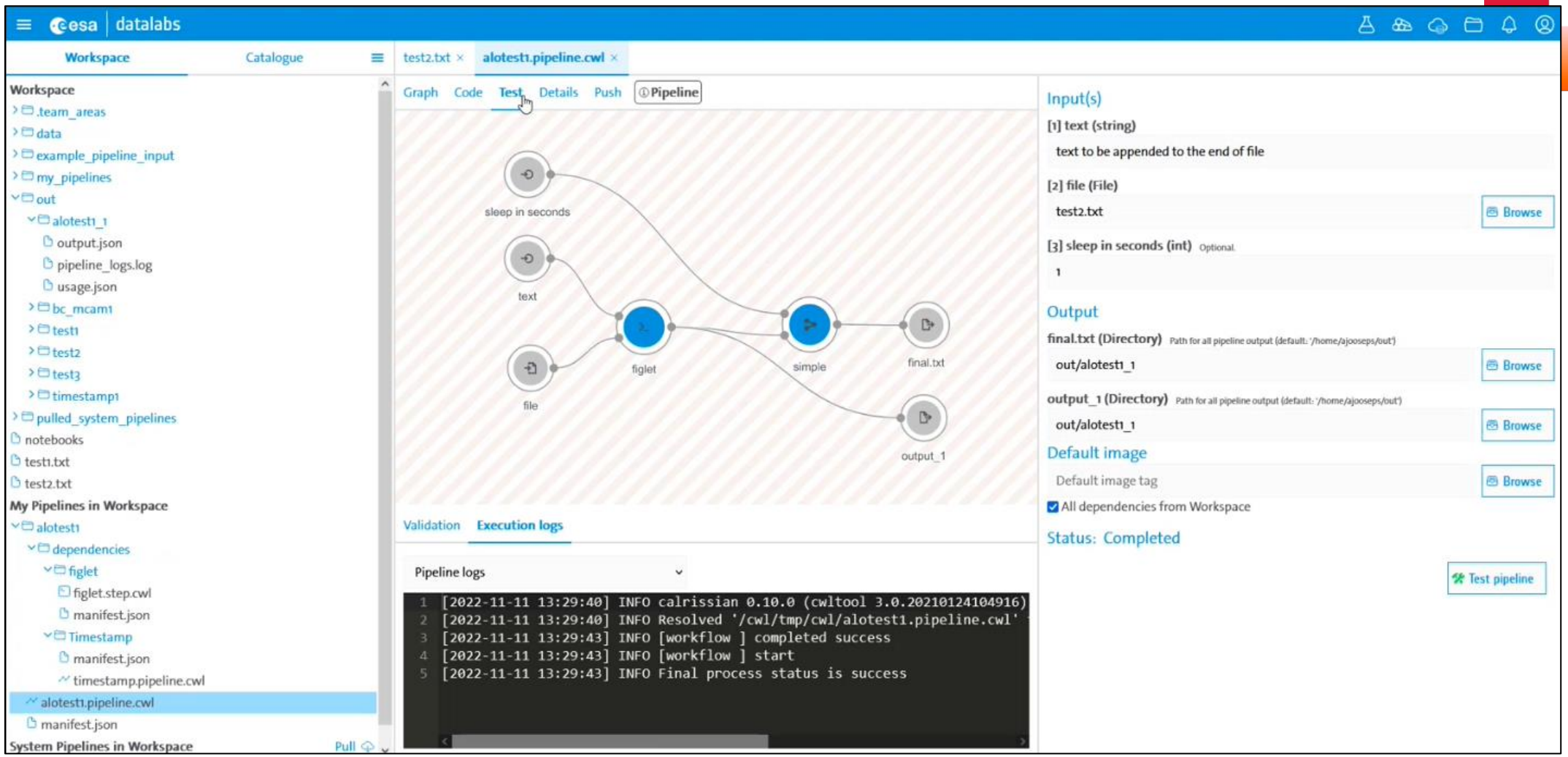




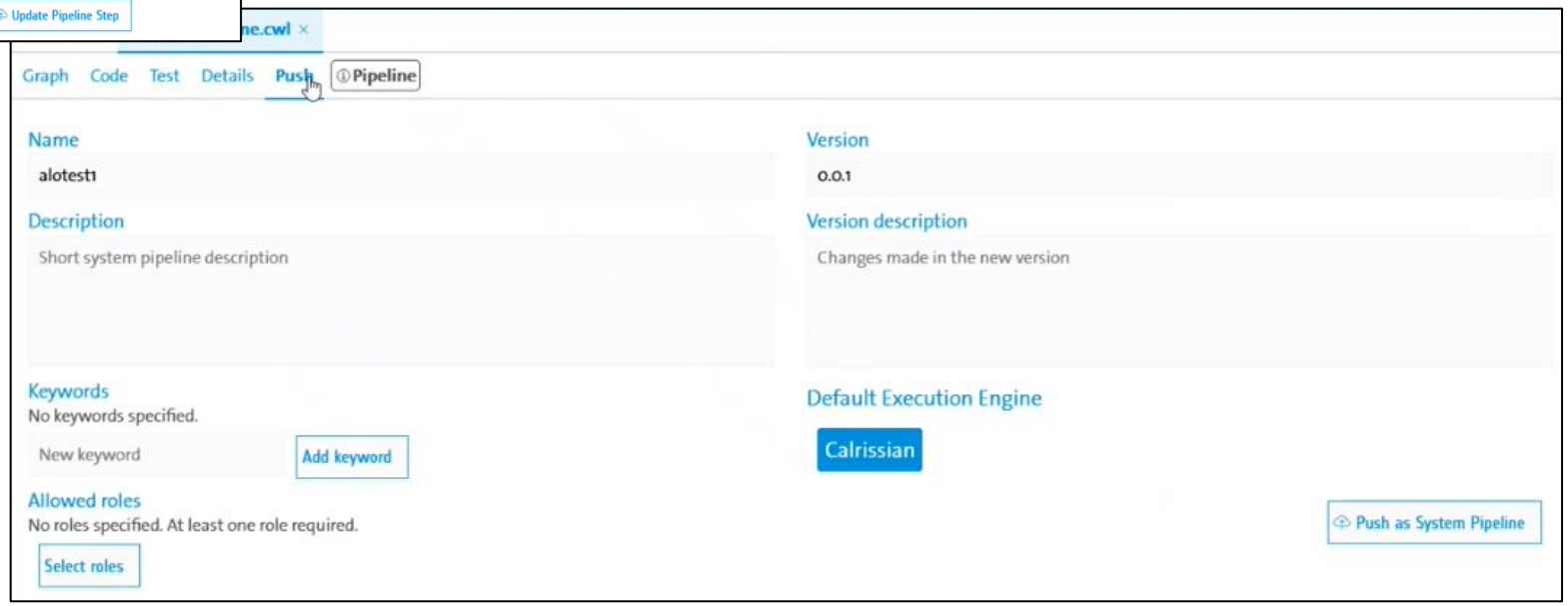
Pipeline Catalogue – Find, Pull, Modify



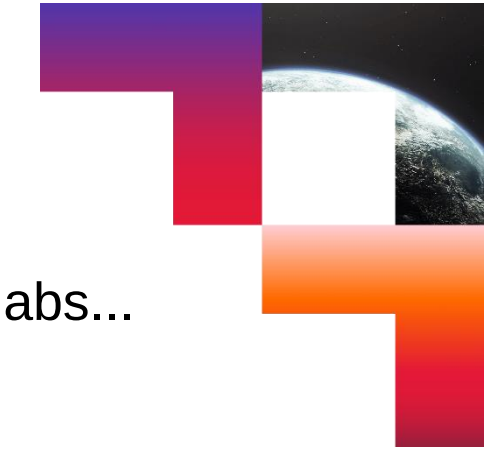
Pipeline Editor – Edit, Code, Test, Details



Pipeline Catalogue – Access, Version, Keywords, Push



Release plan



Version 1.0 of pipelines module available for public use in ESA datalabs...

v1.0 January 2023

and quarterly releases after that

Benefits of using ESA Datalabs Pipelines

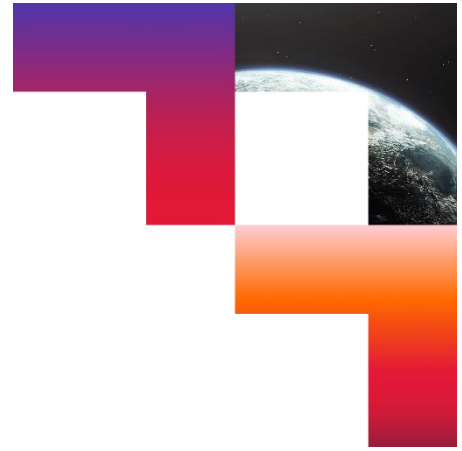
1. Easier access to data which is close to the processing location
2. Easier to share with team(s) and (if needed) also with wider researcher community
3. Easier to define input-output changes
4. Easier to schedule, monitor, stop etc runs
5. Easier to get and modify notifications about run events
6. Easier to set-up and run your code (installation of environment lower levels and dependencies are taken care of by professional IT)
7. Easier to parallelize the runs

Ideas for additional features for ESA Datalabs pipelines

1. Easier uploading for step code or step configuration in ESA Datalabs pipelines.
2. Clearer way to define whose intellectual property the pipeline consists of and how to refer to it when it was used in your work.
3. Ordering the launching of several pipeline instances in parallel with access to a lot of memory and high performance computing power.
4. Ability to take input from and save output to external cloud data sources (Google Drive, Dropbox, Microsoft Onedrive).
5. More intelligent and dynamic input and output location selection using some pattern.
6. Automatic run when new data is available (e.g. listening to the folder).
7. Ability to obtain input data from databases and save output data to databases.
8. Step code log event generated e-mail notification.
9. Conditional (i.e. branching) pipeline definition.

Your feature wishes and priorities are very welcome for planning the next releases.

Pipelines Hands-On session 25.11.22 11:30-12:30



1. Please register **TODAY** to become a user of the pipelines demo environment for tomorrow’s hands-on session – the background processing of registration is slow.
2. Connect to wireless network „ESA“
3. URL to start registration: https://datalabs-pipeman.situk8spro.ebe.lan
4. NB! this is a separate environment from the other hands-on sessions, so separate registration is necessary.

Your feedback, questions and feature requests about pipelines in ESA Datalabs are welcome!

Write to alo.joosepson@cgi.com

Thank you! Gracias!

cgi.com