

CIT

Lessons from the Deployment of a Production iRODS Auditing Module

JGM 2026



Authors

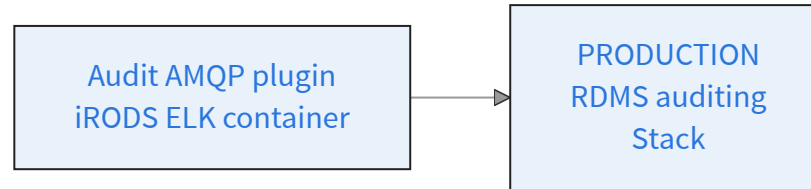
- Venustiano Soancatl Aguilar
- Simona Stoica
- Luca Le Preux
- Ger Strikwerda
- Harm Vos
- John Mc Farland
- Burcu Beygu Koopmans
- Andrey Tsyganov

Motivation

Researchers want to see what happened to their own data:

- Create
- Open
- Move/rename
- Add/remove/modify metadata

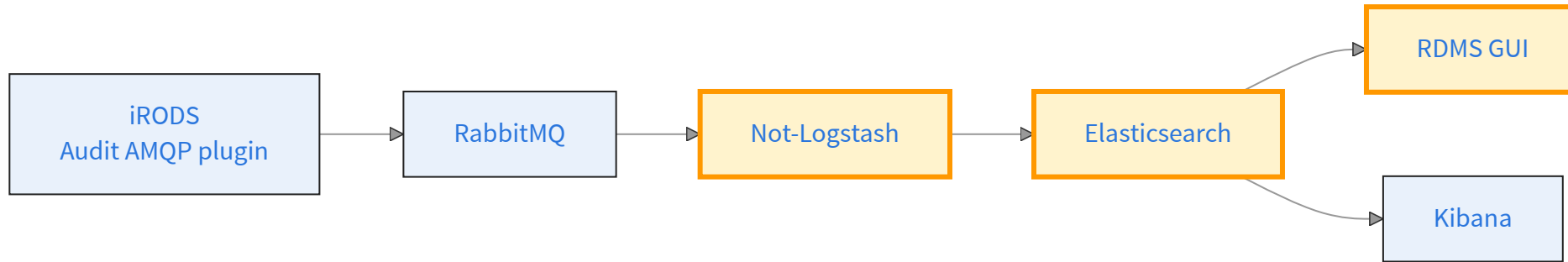
From POC to Production



Key engineering concerns:

- event selection and filtering
- track log activity on iRODS objects

Reference Architecture



Design goal: keep the operational pipeline reliable while exposing meaningful audit trails to end users.

First Setup

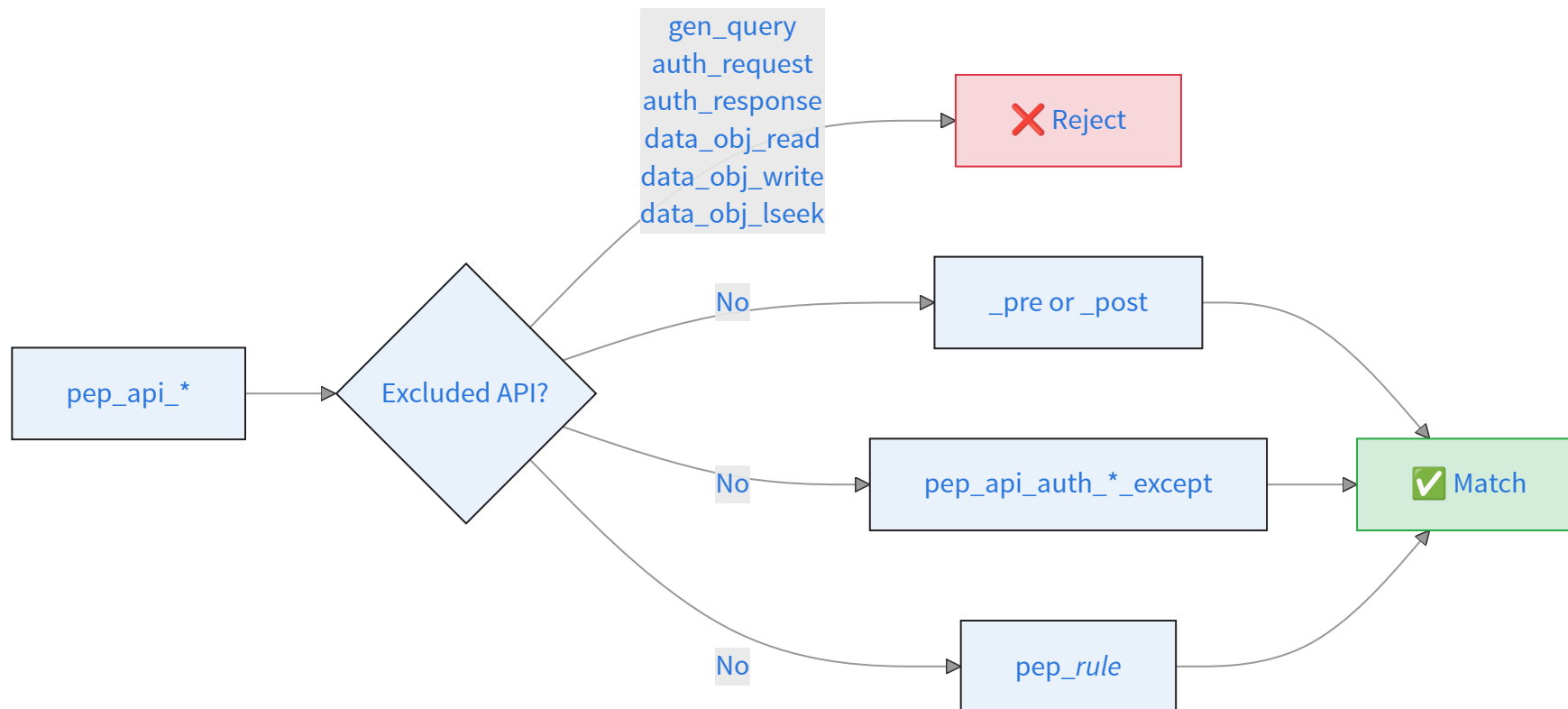
- 1 elasticsearch instance
- 1 single Index
- ~~Broke after 6-7 months and a 700 GB index~~

Second Setup

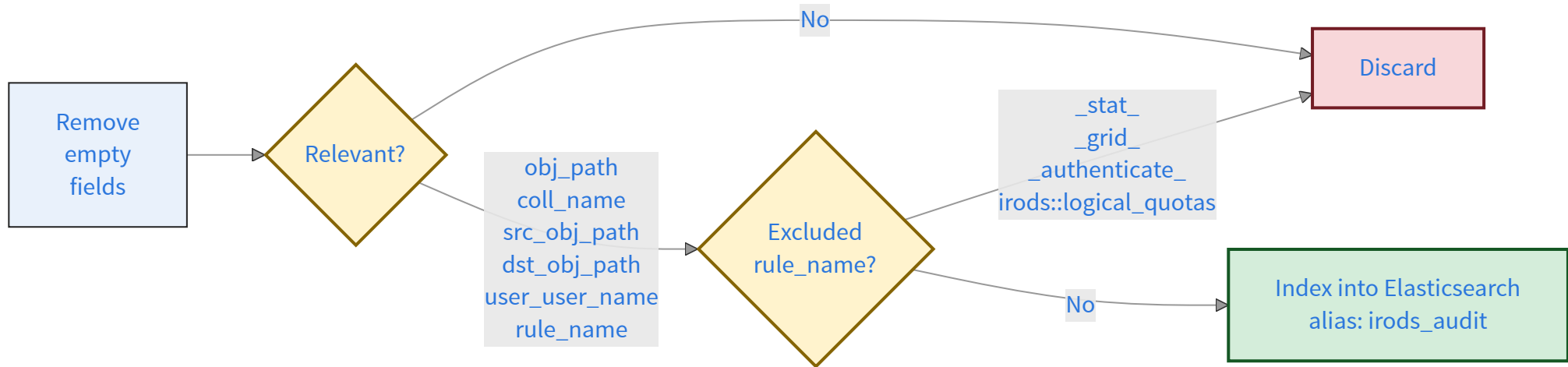
- Hot-Warm-Cold architecture
- Filter based on a regular expression in the iRODS settings
- ~~Filter in Logstash (ruby)~~
- Filter in Not-Logstash (python)
- Elasticsearch deployed via ~~helm charts~~

Filter in the iRODS settings

```
1 "pep_regex_to_match":  
2 "(pep_api_(?!gen_query|auth_(response|request)|data_obj_(lseek|read|write))  
3 |pep_api_auth_.*_except  
4 |pep_.*rule.*)"
```



Not logstash filtering



By applying the filters we reduced ~80% of the documents ingested into elasticsearch.

Capacity Planning

- User model projection
- Data model projection

Population Growth Model

$$U(t) = U_0(1 + g_u)^t$$

- U_t is the number of users at year t .
- U_{t-1} is the number of users at year $t - 1$.
- g_u is the annual user growth rate.

$$g_u = \left(\frac{\text{Ending Value}}{\text{Beginning Value}} \right)^{\frac{1}{n}} - 1 = \left(\frac{150}{22} \right)^{\frac{1}{5}} - 1$$

g_u is estimated using the compound annual growth rate (CAGR) formula.

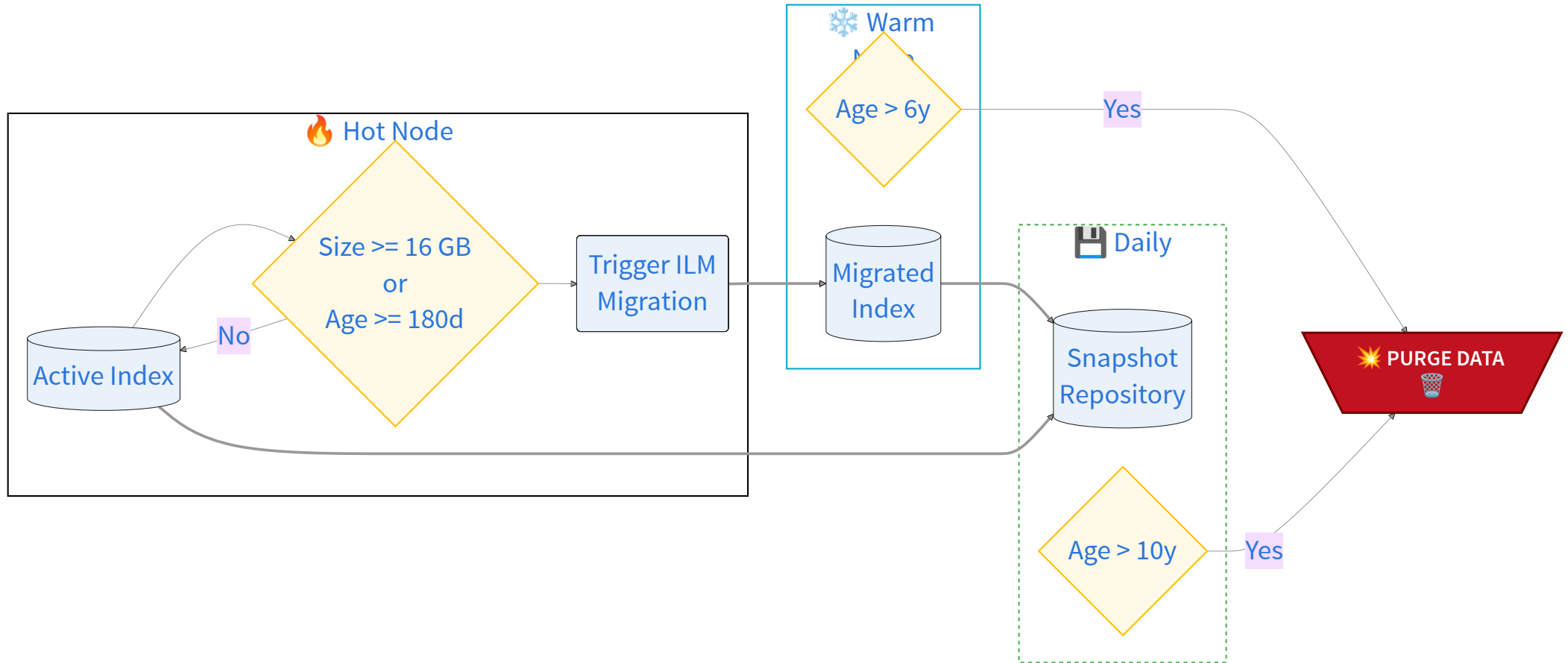
Data Volume Projection

$$D_t = D_{t-1} \times (1 + g_d) + (U_t \times d_u)$$

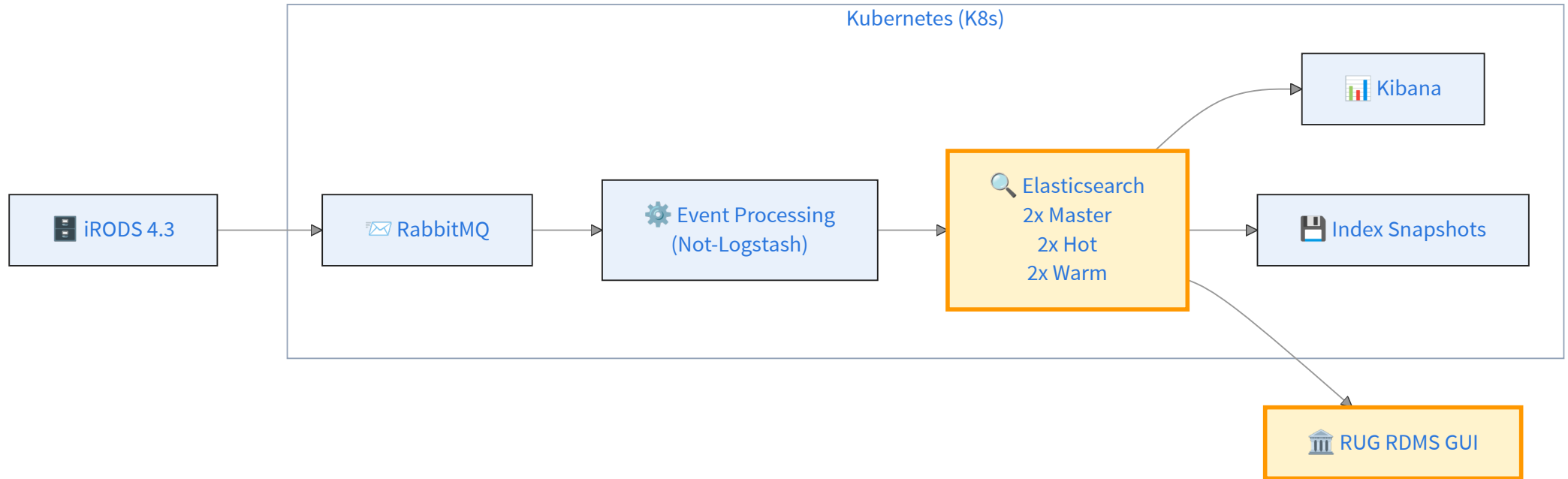
- D_t : Data volume at year t .
- D_{t-1} : Baseline data at year $t - 1$.
- g_d : Annual data growth rate.
- U_t : Total number of active users at year t .
- d_u : Average data generated per user annually.

The left side handles exponential background system growth (logs, backups), while the right side handles linear user-driven growth scale.

Storage Lifecycle Management



The elastic stack



The RDMS GUI

The screenshot displays the RDMS GUI interface, which is divided into three main sections: a sidebar, a central file explorer, and a right-hand panel.

Sidebar: The sidebar on the left contains a navigation menu with icons for home, search, and other functions. Below the icons, the text "My folder" is displayed, followed by the email address "soancatl.aguilar@rug.nl". A search bar labeled "Filter by name.." is present. Below the search bar, there are several expandable sections: "Starred", "Projects", "Team drives", "Shared", "Bin", "My folder", "collection_rename", "test", and "test_history". The "test_history" section is currently selected and highlighted.

Central File Explorer: The central area shows a file explorer view for the "test_history" folder. The path is "/devrugZone/home/v.soancatl.aguilar@rug.nl/test/test_history". A search bar labeled "Filter data.." is present. Below the search bar, a table lists the files in the folder:

Name
ugm-2026.qmd

The file "ugm-2026.qmd" is highlighted. A tooltip indicates "shown 1 from 1 files/folders".

Right-hand Panel: The right-hand panel displays the details for the selected file "ugm-2026.qmd". It includes a header with the file name and icons for refresh and close. Below the header, there are several icons for file operations. The main content area shows a list of activities:

action	timestamp
remove metadata by v.soancatl.aguilar@rug.nl	2026-06-18 11:33:58
set metadata by v.soancatl.aguilar@rug.nl	2026-06-18 11:29:19
set metadata by v.soancatl.aguilar@rug.nl	2026-06-18 11:24:56

Below the activity log, there is a section titled "Activity Log" with a table of metadata:

action	timestamp
open by v.soancatl.aguilar@rug.nl	2026-06-18 10:54:08
create by v.soancatl.aguilar@rug.nl	2026-06-18 10:54:03

Path Evolution History

```
1 .../test/test_history/ugm-2026.qmd
2 .../test/archive/ugm-2026.qmd      # Fold
3 .../production/archive/ugm-2026.qmd # Path
4 .../production/archive/ugm-2026-final.qmd
```

A recursive algorithm was implemented to find the evolution history.

ugm_2026-final.qmd activity log

Activity Log	
action	rename
timestamp	1782141115857
username	v.soancatl.aguilar@rug.nl
source	/devrugZone/home/v.soancatl.aguilar@rug.nl/production/archive/ugm-2026.qmd
destination	/devrugZone/home/v.soancatl.aguilar@rug.nl/production/archive/ugm-2026-final.qmd

rename by v.soancatl.aguilar@rug.nl	2026-06-22 12:59:42
rename by v.soancatl.aguilar@rug.nl	2026-06-22 11:49:57
open by v.soancatl.aguilar@rug.nl	2026-06-18 11:46:50
remove metadata by v.soancatl.aguilar@rug.nl	2026-06-18 11:33:58
set metadata by v.soancatl.aguilar@rug.nl	2026-06-18 11:29:19
set metadata by v.soancatl.aguilar@rug.nl	2026-06-18 11:24:56
open by v.soancatl.aguilar@rug.nl	2026-06-18 10:54:08
create by v.soancatl.aguilar@rug.nl	2026-06-18 10:54:03

Activity Log	
action	create
timestamp	1781780043813
username	v.soancatl.aguilar@rug.nl
path	/devrugZone/home/v.soancatl.aguilar@rug.nl/test/test_history/ugm-2026.qmd

Operational constraints

The 2025 Bitnami Paywall

- ~~Elasticsearch Helm charts~~
- ~~RabbitMQ Helm charts~~
-  ECK operator
-  RabbitMQ operator ?

RabbitMQ connection

- ~~guest/guest~~ can not connect remotely in RabbitMQ 3.9+
- Plain credentials are disabled by default in Proton
 - Test compiling the irods audit plugin

```
1 // use it only for closed environments
2 sasl_allow_insecure_mechs(true);
3 sasl_allowed_mechs("PLAIN");
```

Extra test

```
1 // after each proton::message initialization
2 msg.durable(true);
```

Note: The [irods/irods_development_environment](#) repository is really handy to quickly compile the `iRODS audit plugin`.

Next Steps

Refining event and filtering rules

Triggered **peps** during 15 minutes

rule_name		
Top values		
pep_api_auth_plugin_request_post	23.0%	⊕ ⊖
pep_api_auth_plugin_request_pre	23.0%	⊕ ⊖
pep_api_get_misc_svr_info_post	14.0%	⊕ ⊖
pep_api_get_misc_svr_info_pre	14.0%	⊕ ⊖
pep_api_specific_query_post	13.1%	⊕ ⊖
pep_api_specific_query_pre	13.1%	⊕ ⊖
Calculated from 222 records.		

Can we exclude it?

api_index	700
auth_scheme	native
client_addr	129.125.77.28
connect_count	1
ERROR	[MiscSvrInfo**] not supported
hostname	store-dev.data.rug.nl
option	imiscsvrinfo
pid	1,788,310
proxy_auth_info_auth_flag	0
proxy_auth_info_flag	0
proxy_auth_info_ppid	0
proxy_rods_zone	devrugZone
proxy_sys_uid	0
proxy_user_name	irods
rule_name	pep_api_get_misc_svr_info_post
socket	0
status	0

Other Steps

- Node monitoring and alerting
- Recovering deleted indices from S3 snapshots
- Upgrade RabbitMQ
- ...

Takeaways

Key lessons for building a production-ready auditing service:

- **Don't start from scratch:** leverage existing iRODS plugins and the elastic stack.
- **Keep in mind the 80-20 rule:** Focus on the vital 20% that drives 80% of the impact.
- **Be open to change:** External policies evolve rapidly, and staying adaptable is crucial.