Data: The Final Frontier

John Constable
Informatics Support Group,
Informatics Digital Solutions
Wellcome Sanger Institute

These are the voyages of the Informatics Digital Solutions team at Sanger.

Its (fourteen so far) year iRODS mission:

To migrate old data.

To seek out new features.

To boldly go where no iRODS Zone has gone before!

Quick Recap

Who am I

Principal System Administrator in the team that looks after the HPC and OpenStack environments at Sanger (which includes Lustre and iRODS storage systems).

Managing iRODS at Sanger since 2014.

Shaver of Yaks.

What is Sanger

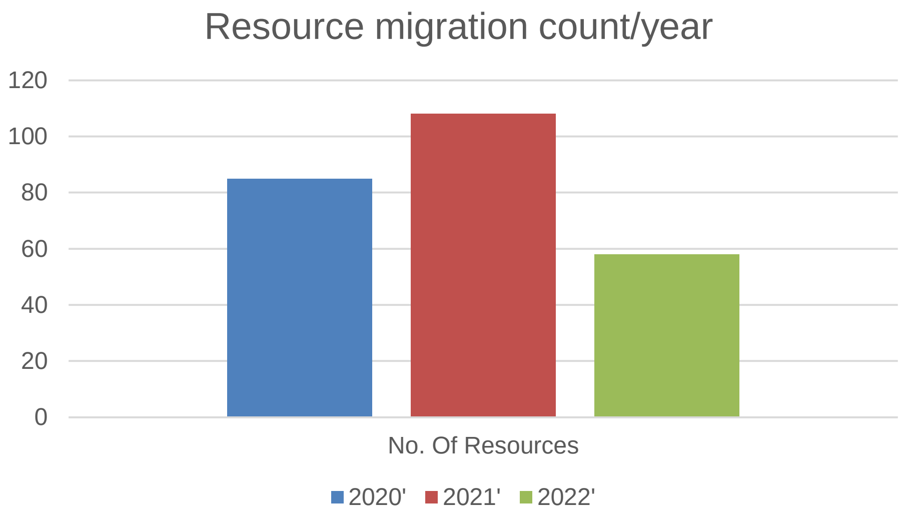
The Wellcome Sanger Institute is a world leader in genome research that delivers insights into human, evolutionary and pathogen biology.

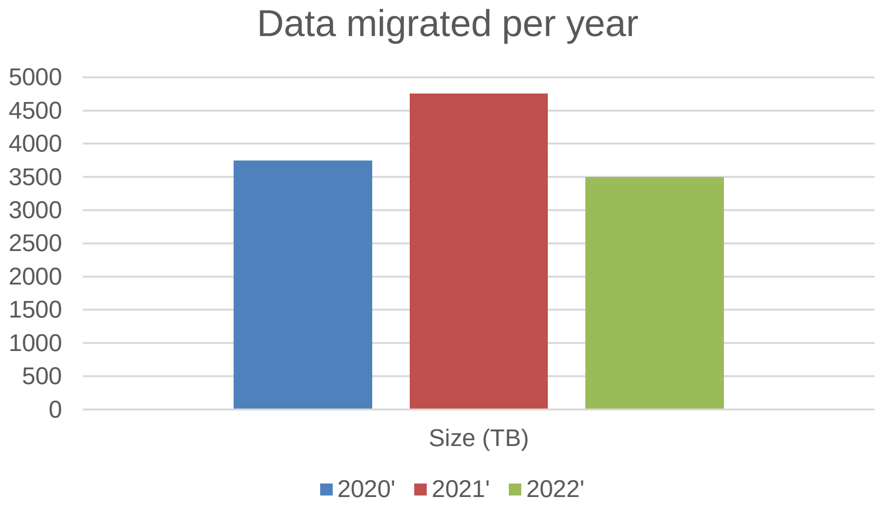
https://www.sanger.ac.uk/about/our-vision/

We have over a hundred iRODS servers, mostly storage, around 50PB (most is replicated so half that usable).
Running 4.2.7 and Ubuntu 18.04

Migrating 8PB of data

Migrating 8PB of data


Resource migration count/year


Data migrated per year

wellcome sanger institute

Migrating 8PB of data – the hero

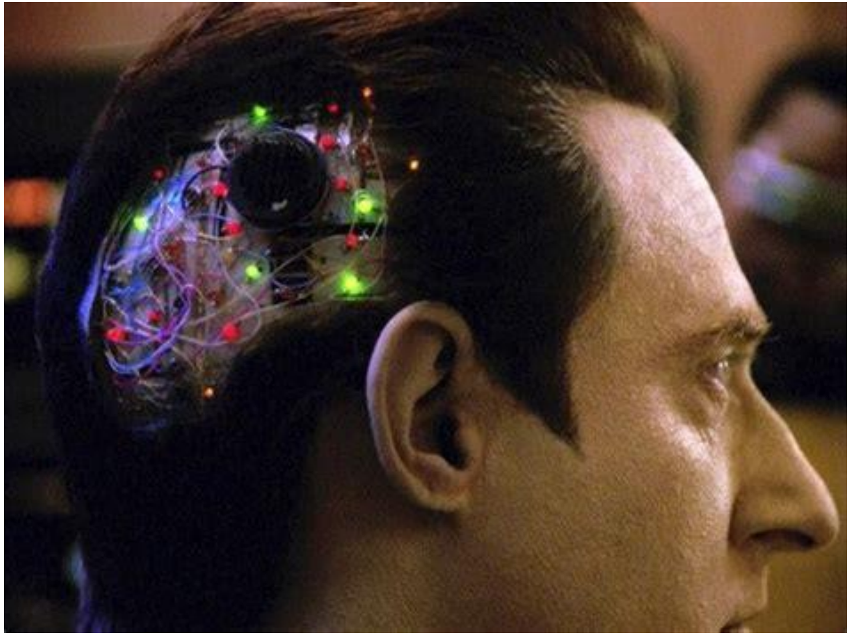
Vijay Arangarajan
60 servers
decommissioned!

Migrating 8PB of data – automation saved us

- There is no current tool to migrate a resource to a new resource, whether in a composite tree or not.
- Hands up if you'd use one if it was built in?
- In future, would we do it with the tiering plugin? Maybe…
- Look to last years talk about how/why we did it in more detail.

Switching to PostgreSQL

Switching to PostgreSQL



- Show of hands – PostgreSQL, MySQL, Oracle?
- Previously Oracle, now PostgreSQL on all Zones.
- Better consortium take up of PostgreSQL – few using Oracle.
- Also cost savings!
- ~~300 350~~ 400 million items, roughly similar amounts of metadata
- Upgrade was easy, but quite disruptive due to quantity of data to export (800GB), convert, copy and then reload.
- Expert DBA team were amazing
    - ran multiple conversion dry ruins on copies of database prior to the real thing
    - worked with us to run performance and load tests.
- Then we found out about `iadmin rum` – might have saved us 20% of the data!

Migrating to Single Replicas using the Tiering Plugin

Tiering Plugin

- How many attendees using the tiering plugin?
- Normally store two replicas.
- Some data sent to the EGA, so drop to one replica for those objects.
- Aim is to tag items we want to move with metadata, and the plugin moves it from a composite tree with a replica pass-thru to a composite tree with just random ones.
- Ideally tag the lot and let the plugin move items over time.

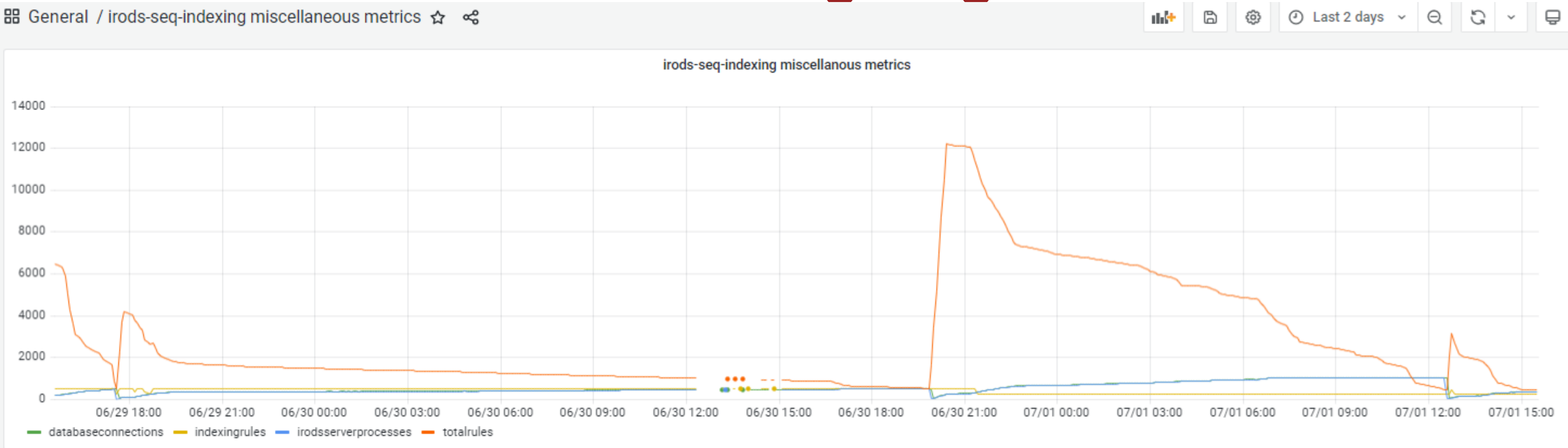
Important to be sure we get the right Data

✗ ✗ ✓

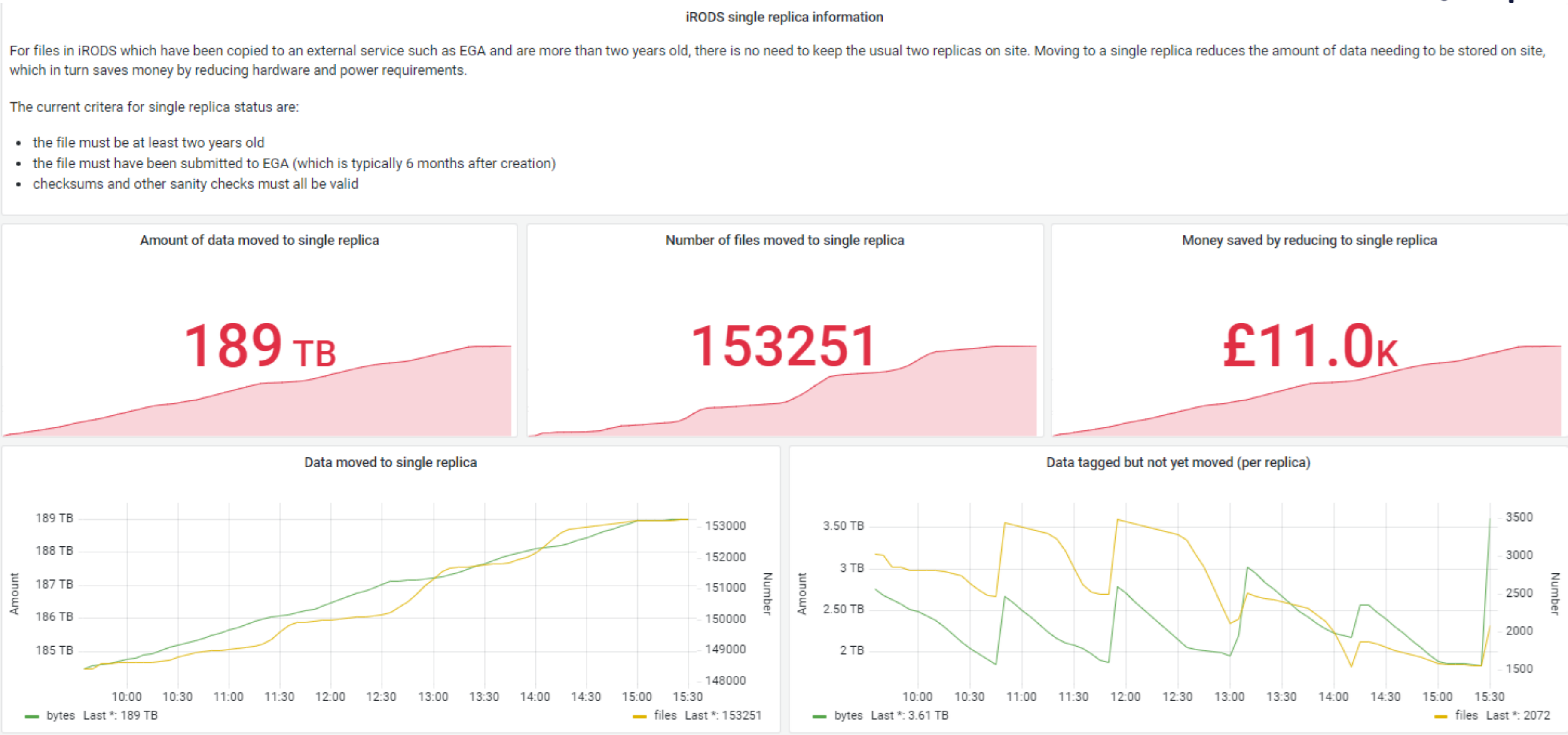
Tiering Plugin

- Not always easy to access right documentation – we're on 4.2.7, so README was for 4.2.11 when we set it up.
- Default is time based, so files without that metadata tag wont move. This means we need to tag an item to move with multiple items of metadata
- 4.2.7 doesn't throttle max no of concurrent rules (known bug bug), so we had to experiment with the in built throttling for the plugin, which worked, sort of.
- 4.2.7 doesn't drop connections to the database until all the items in that processes set of items found by the violating objects query have moved.
- This means we need to manage the queue manually to stop running the database out of connections, and we cant just tag all the items.
- rodsLog is very noisy with LOG_NOTICE enabled, but that's the only(?) way to get a report of what's happening.
- Basically we need to upgrade to 4.2.11, but that's non-trivial to arrange!
- Side benefit is addition of access time metadata allows us to show whats been used recently!

Single Replica Migrations aka Tiering Plugin



Single Replica Migrations



iRODS single replica information

For files in iRODS which have been copied to an external service such as EGA and are more than two years old, there is no need to keep the usual two replicas on site. Moving to a single replica reduces the amount of data needing to be stored on site, which in turn saves money by reducing hardware and power requirements.

The current critera for single replica status are:

- the file must be at least two years old
- the file must have been submitted to EGA (which is typically 6 months after creation)
- checksums and other sanity checks must all be valid

Amount of data moved to single replica

189 TB

Number of files moved to single replica

153251

Money saved by reducing to single replica

£11.0K

Data moved to single replica

bytes Last *: 189 TB
files Last *: 153251

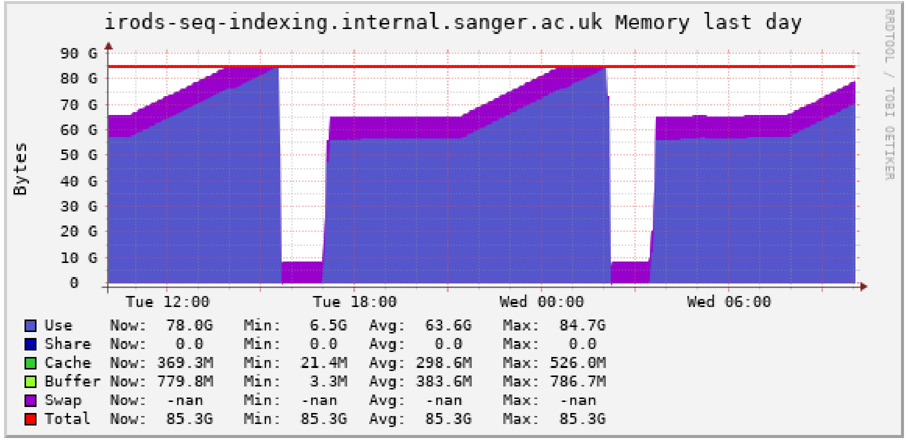
Data tagged but not yet moved (per replica)

bytes Last *: 3.61 TB
files Last *: 2072

Improving the searching of 400 million items of metadata

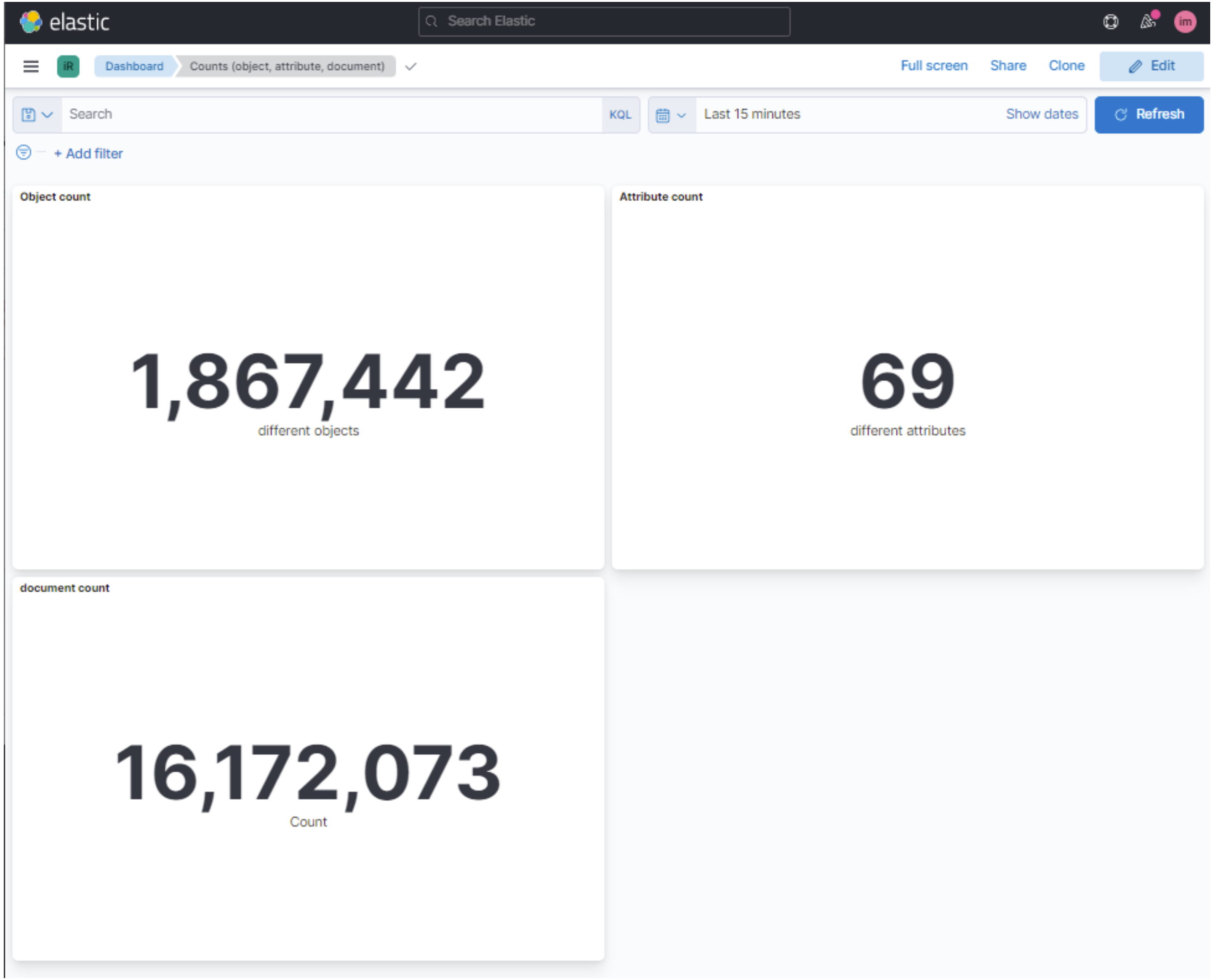
wellcome sanger institute

Improving the searching of 400 million items of metadata



	Now:	Min:	Avg:	Max:
🟦 Use	78.0G	6.5G	63.6G	84.7G
🟦 Share	0.0	0.0	0.0	0.0
🟩 Cache	369.3M	21.4M	298.6M	526.0M
🟩 Buffer	779.8M	3.3M	383.6M	786.7M
🟪 Swap	-nan	-nan	-nan	-nan
🟥 Total	85.3G	85.3G	85.3G	85.3G

- Adding metadata at a fantastic rate – *at least* 50 million a year.
- This rate is increasing, despite measures such as adding metadata to collections rather than every data object.
- Switching from Oracle to PostgreSQL (see later) didn't help as with Oracle we can 'pin' queries that select a non-optimal ordering.
- Trying the Indexing Plugin (ElasticSearch).
- Currently it runs the server out of memory at least twice a day.
  - I suspect its trying to build all the rules in one go.
  - Pretty much the largest VM we can run the delay server on, ATM.
  - Not actually updating the ElasticSearch Index, as far as I can tell.
- A work in progress? Very promising if it does happen though!
- If your end users don't want to use imeta/iquest they surely wont want to use curl/JSON search! We wrote iimeta to help.
- How many attendees using the indexing plugin?



Counts (object, attribute, document)

Last 15 minutes

Object count

1,867,442

different objects

Attribute count

69

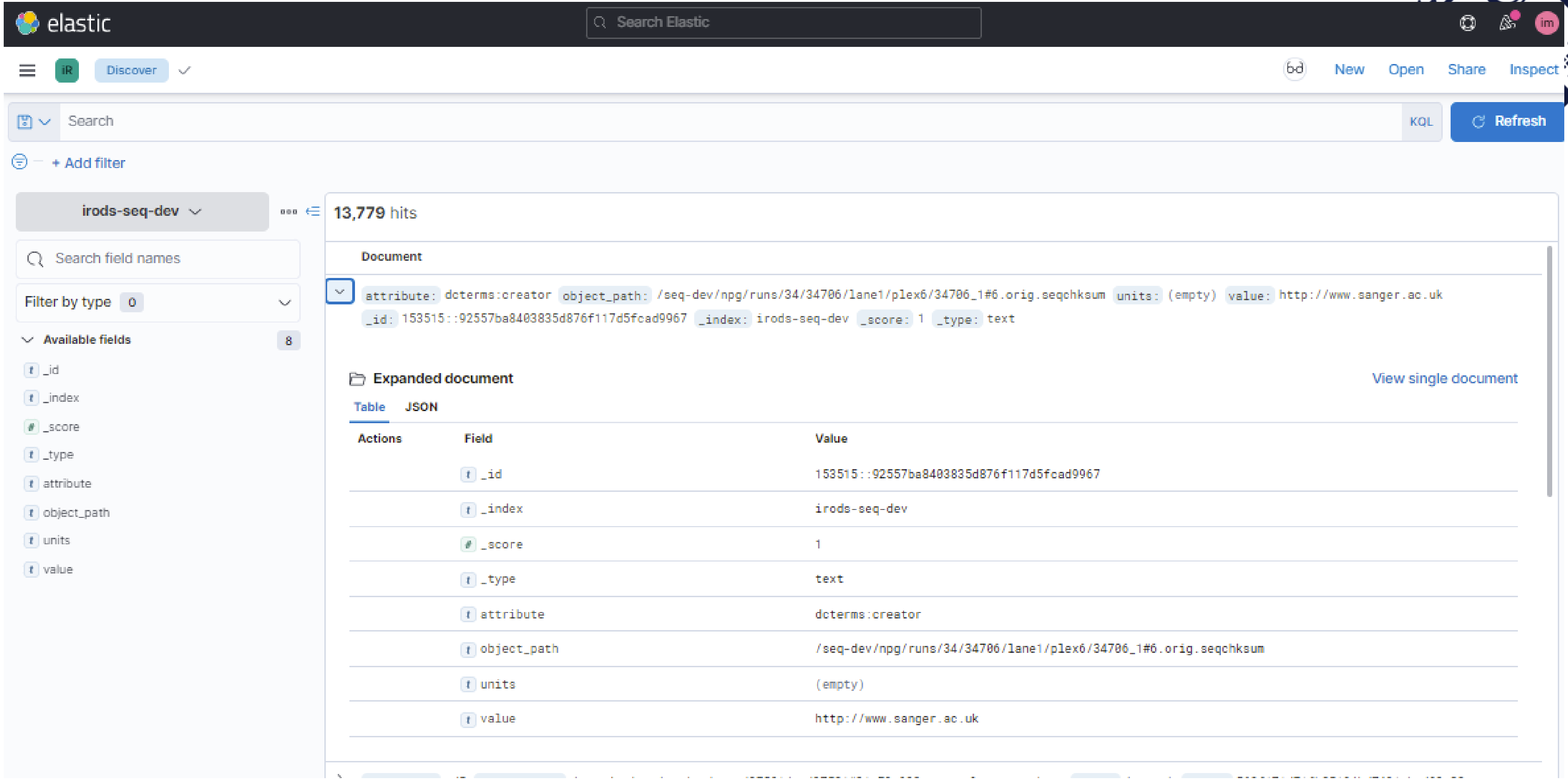
different attributes

document count

16,172,073

Count

wellcome sanger institute



13,779 hits

Document

```
attribute: dcterms:creator object_path: /seq-dev/npg/runs/34/34706/lane1/plex6/34706_1#6.orig.seqchksum units: (empty) value: http://www.sanger.ac.uk
_id: 153515::92557ba8403835d876f117d5fcad9967 _index: irods-seq-dev _score: 1 _type: text
```

Expanded document

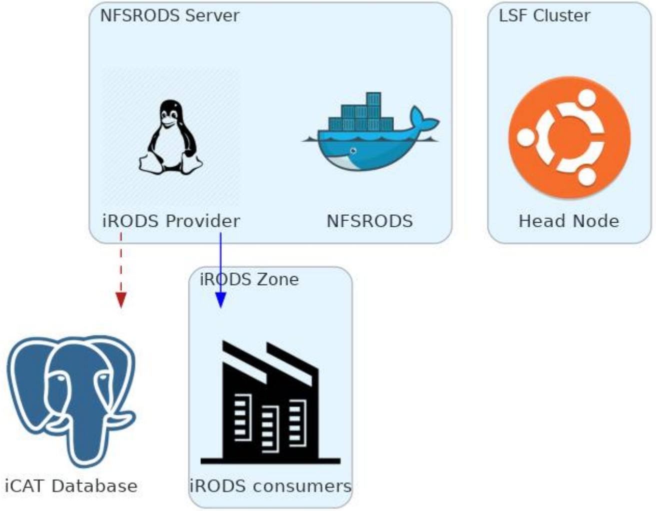
View single document

Table JSON

Actions	Field	Value
	_id	153515::92557ba8403835d876f117d5fcad9967
	_index	irods-seq-dev
	_score	1
	_type	text
	attribute	dcterms:creator
	object_path	/seq-dev/npg/runs/34/34706/lane1/plex6/34706_1#6.orig.seqchksum
	units	(empty)
	value	http://www.sanger.ac.uk

Deploying NFSRODS

Deploying NFSRODS



Sanger NFSRODS deployment

- Deployed to a federated zone.
- Zone has to have local users which get the permissions (user#local) *and* federated users (user#FederationMaster) as NFSRODS only maps local users to NFS UID's.
- The ACL's and users have to be kept in sync!

One VM with both Provider and NFSRODS:

- scalability - can have N of them, and potentially round-robin NFS mount.
- Performance – shortest hop between NFSRODS and Catalog/Provider

Deploying NFSRODS

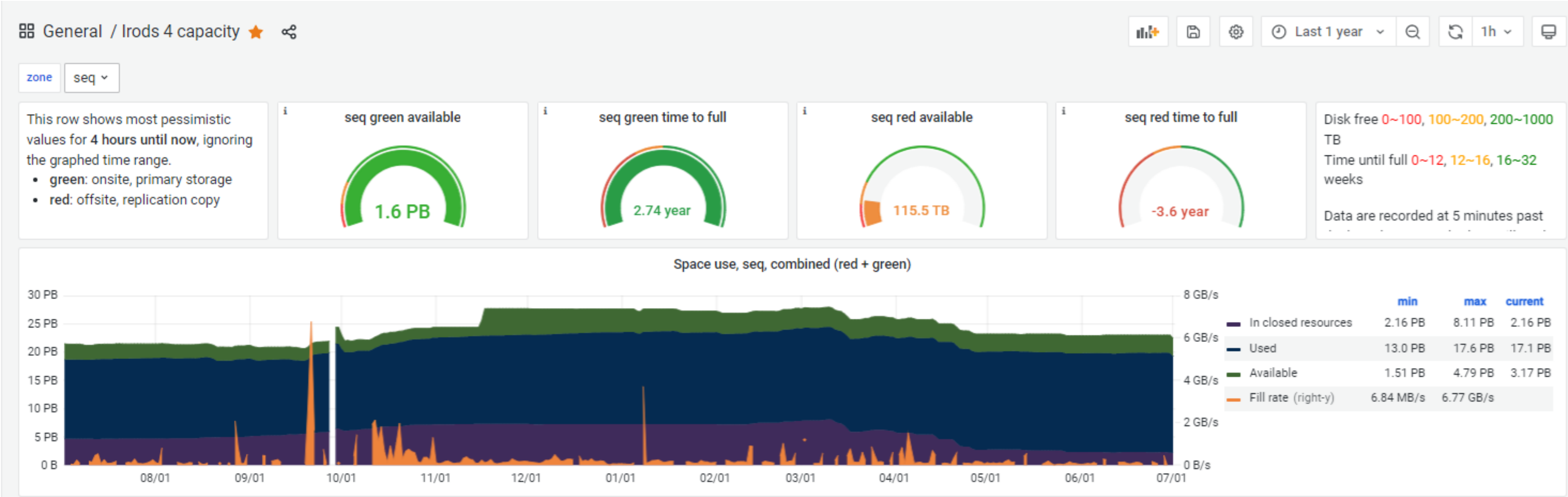
Performance

- Initially shocking (did too many stat calls and lookups)
- Then poor - slow, single threaded.
- Some issues around collection with 64k files (I know, I know) – might be resolved now?
- Latest release should be much better - multi threaded (but not tested yet)

Hands up - who else is using this?

Deploying the usual few petabytes of storage

Deploying the usual few PB of storage



General / Irods 4 capacity

zone seq

This row shows most pessimistic values for **4 hours until now**, ignoring the graphed time range.

- **green**: onsite, primary storage
- **red**: offsite, replication copy

Disk free 0~100, 100~200, 200~1000 TB
Time until full 0~12, 12~16, 16~32 weeks

Data are recorded at 5 minutes past

	min	max	current
In closed resources	2.16 PB	8.11 PB	2.16 PB
Used	13.0 PB	17.6 PB	17.1 PB
Available	1.51 PB	4.79 PB	3.17 PB
Fill rate (right-y)	6.84 MB/s	6.77 GB/s	

Deploying the usual few PB of storage

We use DDN, Supermicro and Dell servers.

Oh, the supply issues!

- Switches on a 1 YEAR lead time
- Optics months delays
- Supermicro changing ETA for servers, sometimes not even able to commit to one.
- Dell servers arrived months late.
- I bet Scotty didn't have to deal with this!

Shameless Self Promotion Slide!

Twitter: @kript email: jc18@sanger.ac.uk

‘The Resource’ Newsletter - a monthly roundup of developments in iRODS - very geeky, includes github issues!