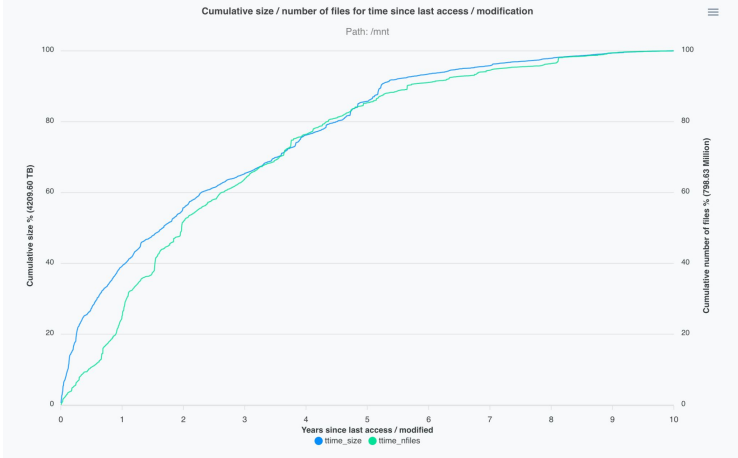
iRODS UGM 2024
Lightning Talk

Emyr James
Head of Scientific IT
CRG Barcelona
emyr.james@crg.eu

mpistat / treemap

- Distributed lstat scan of all inodes in filesystem using libcircle
- atime,mtime, uid, guid, logical size, physical size, file suffix etc.
- Data ingested into clikchouse db for each scan
- Rest api to query data with filters
- Real time aggregation in clickhouse
- Vue.js and python "tui" view
- Early version : https://github.com/CRG-CNAG/mpistat




```
database  : isilon1_2024_05_27
path      : /mnt
size      : 4255.468T
num_files : 815.859M
atime_cost: 1.247M
```

suffix	size	%	+%	num_files	atime_cost
fastq.gz	693.086T	16.3	16.3	1.254M	171.821k
bam	412.168T	9.7	26.0	1.114M	136.026k
tar.gz	203.808T	4.8	30.8	148.788k	35.612k
tif	168.369T	4.0	34.7	20.256M	61.471k
fast5	165.939T	3.9	38.6	42.313M	26.243k
tar	153.456T	3.6	42.2	27.820k	57.293k
fq.gz	102.307T	2.4	44.6	110.633k	22.400k
fastq	100.145T	2.4	47.0	990.393k	28.365k
nd2	77.075T	1.8	48.8	94.529k	33.253k
	75.245T	1.8	50.6	139.585M	31.102k
out.bam	72.482T	1.7	52.3	26.797k	25.737k
raw	70.360T	1.7	53.9	168.851k	20.355k
txt	67.051T	1.6	55.5	38.102M	22.789k
cbcl	60.625T	1.4	56.9	201.184k	3.368k
lif	58.857T	1.4	58.3	40.453k	15.896k
zip	42.078T	1.0	59.3	255.697k	5.642k
tiff	40.301T	0.9	60.2	14.318M	4.659k
vcf.gz	38.786T	0.9	61.1	240.042k	7.343k
tgz	38.301T	0.9	62.0	8.443k	9.340k
bed	35.771T	0.8	62.9	4.489M	7.335k

Upcoming Job Opportunity

- Research Data Management Specialist
- Part of the Scientific IT Team (SIT)
- Bridge between SIT and the Scientists
- RDM Plans, Workflows, Metadata
- Project Management
- Work with IT Technicians to develop systems
- Help to administer systems
- Experience of working with large scale data
- Enthusiasm and evangelism for managing research data the right way

CRG
Centre for Genomic Regulation