DOs and DON'Ts of managing numerous very large databases

A summary of challenges and problems encountered at CERN

Andrzej Nowicki

Voxxed Days CERN 2024

Andrzej Nowicki 🇵🇱

12 years of Oracle DB experience
Database Engineer @ CERN since 2020

andrzejnowicki

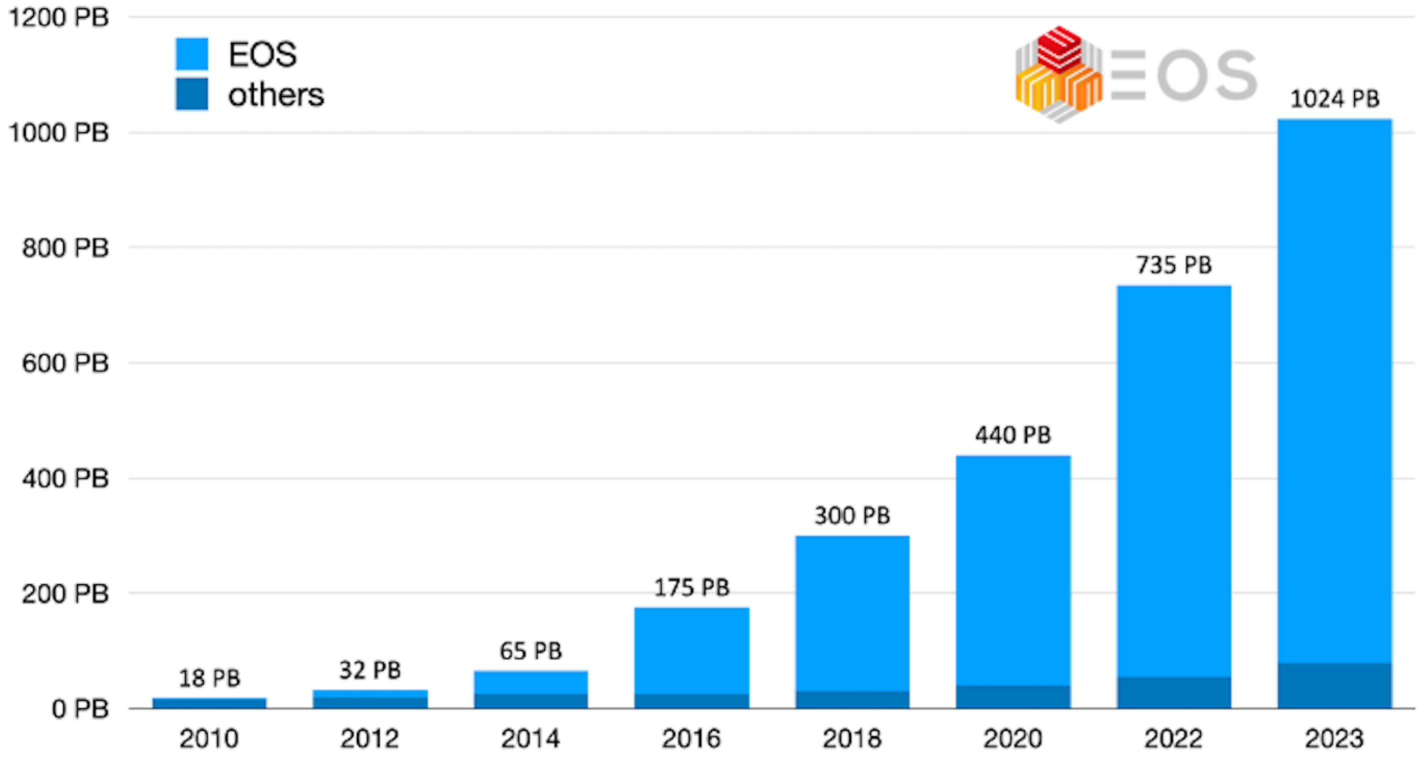
andrzej.nowicki@cern.ch

www.andrzejnowicki.pl

IT @ CERN

An exabyte of disk storage

CERN disk storage capacity passes the threshold of one million terabytes



https://home.cern/news/news/computing/exabyte-disk-storage-cern

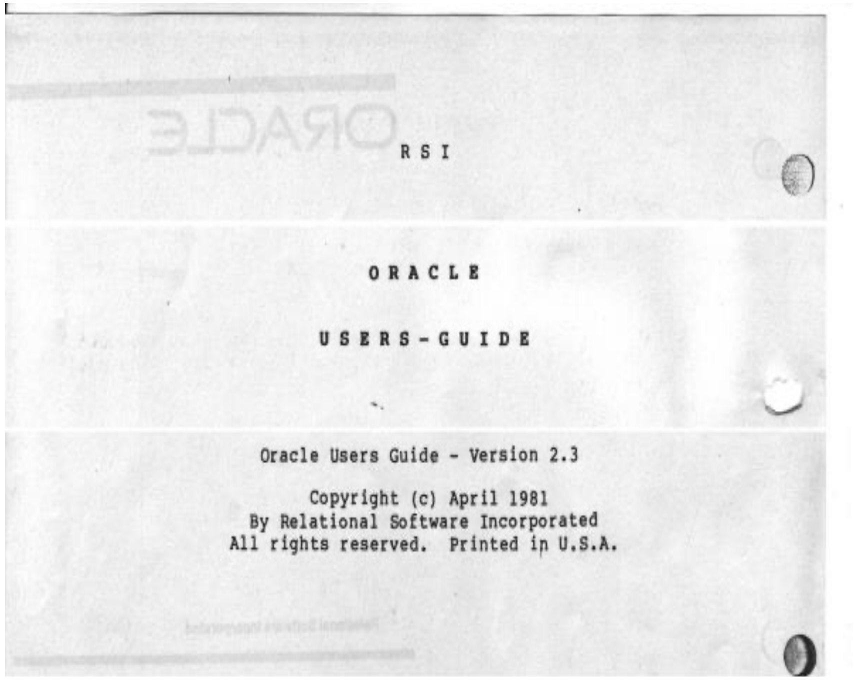
Databases at CERN

Oracle since 1982

- 105 Oracle databases, more than 11.800 Oracle accounts
- RAC, Active Data Guard, GoldenGate, OEM, RMAN, APEX, Cloud…
- Complex environment

Database on Demand (DBoD) since 2011

- ≈600 MySQL, ≈400 PostgreSQL, ≈200 InfluxDB
- Automated backup and recovery services, monitoring, clones, replicas
- HA MySQL clusters (Proxy + primary replica)



RSI

ORACLE

USERS-GUIDE

Oracle Users Guide - Version 2.3

Copyright (c) April 1981
By Relational Software Incorporated
All rights reserved. Printed in U.S.A.

Size of the database environment

	Total size
Oracle	≈5 PB
DBoD (MySQL, PostrgeSQL, InfluxDB)	≈150 TB
Backups	≈3 PB

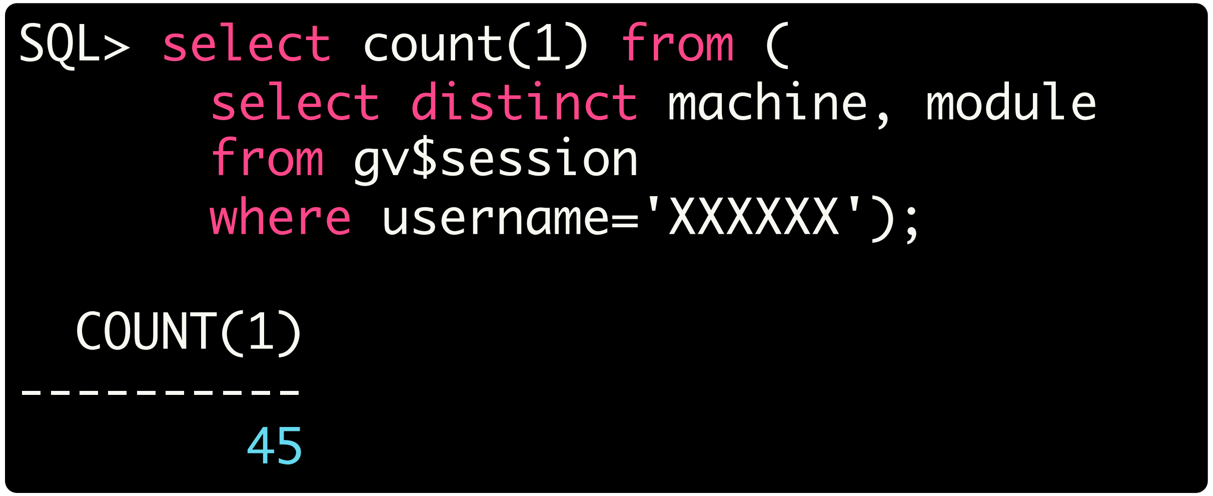
User management

DOs:

- locked schema account
- each application get an account with specific access rights (r/o or r/w)
- provision automatically

DON'Ts:

- OPEN schema account
- password is reused in many applications
- makes an absolute nightmare to manage



```
SQL> select count(1) from (
       select distinct machine, module
       from gv$session
       where username='XXXXXX');

  COUNT(1)
----------
        45
```

Resource management

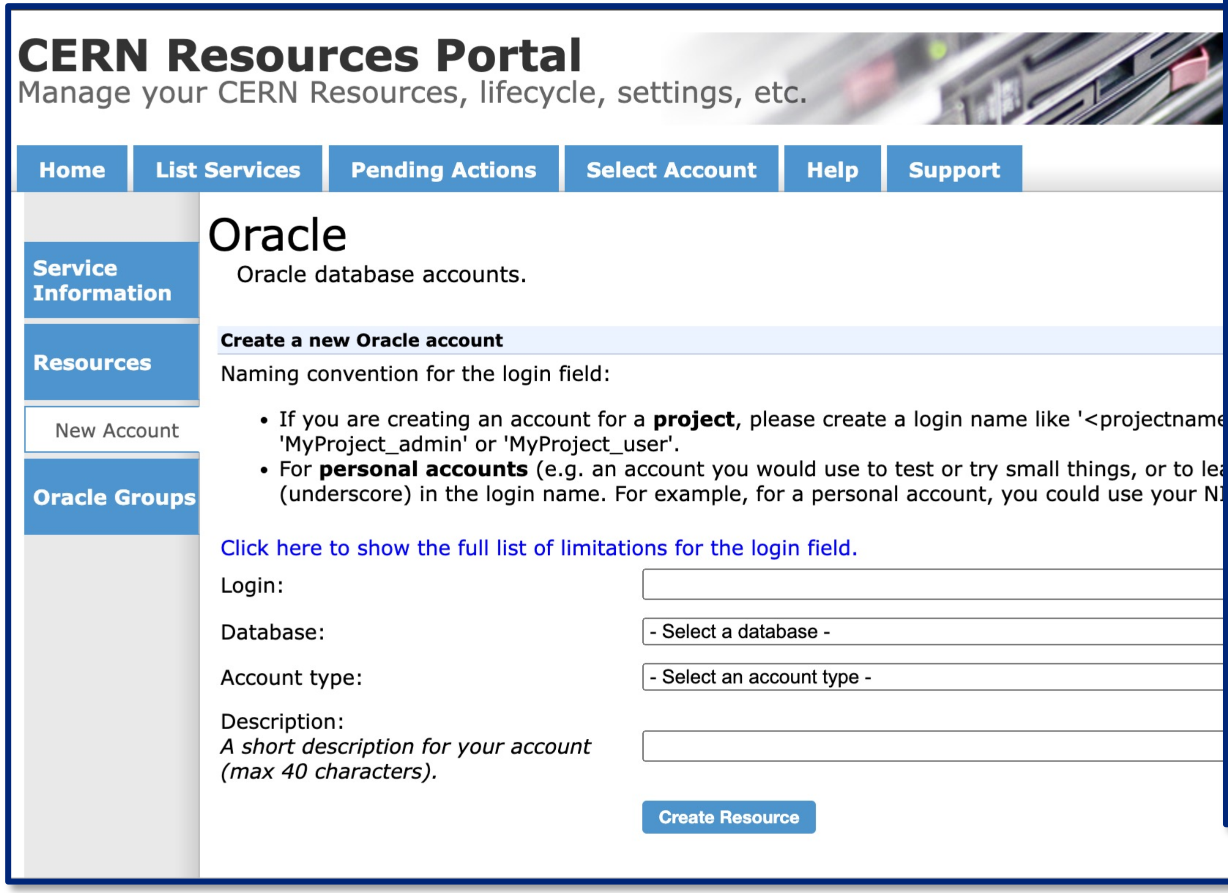
Each resource should have an owner, whom you can contact in case of problems.

We give two options:

- individual owner
- service owner (which might belong to a team, group, etc.)

What do you do when people leave the organisation?

Oracle



CERN Resources Portal

Manage your CERN Resources, lifecycle, settings, etc.

Oracle

Oracle database accounts.

Create a new Oracle account

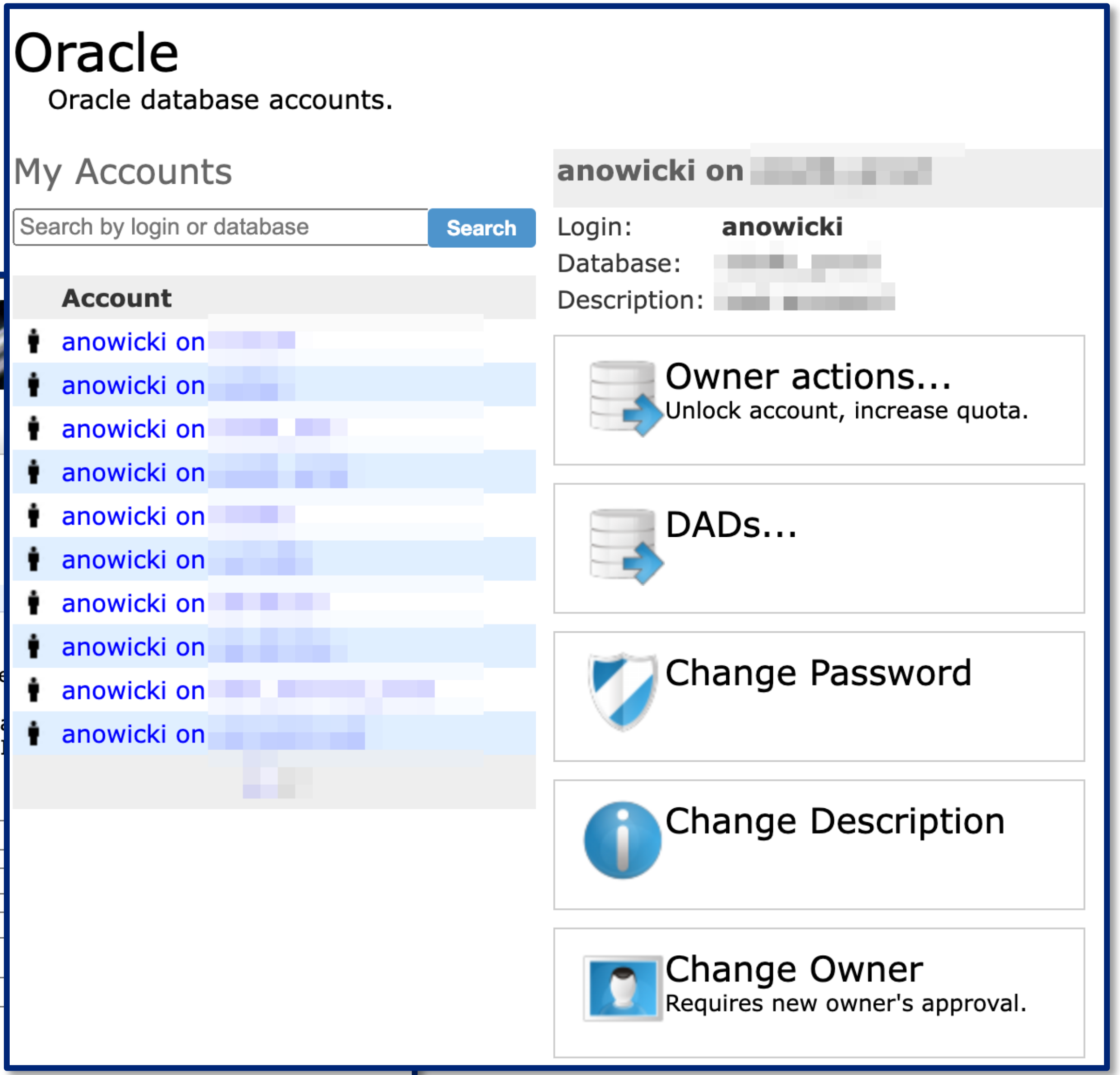
Naming convention for the login field:

- If you are creating an account for a **project**, please create a login name like '<projectname
'MyProject_admin' or 'MyProject_user'.
- For **personal accounts** (e.g. an account you would use to test or try small things, or to lea
(underscore) in the login name. For example, for a personal account, you could use your NI

Click here to show the full list of limitations for the login field.

Login:	
Database:	- Select a database -
Account type:	- Select an account type -
Description: *A short description for your account (max 40 characters).*	

Create Resource



Oracle

Oracle database accounts.

My Accounts

Account
anowicki on [redacted]
anowicki on [redacted]
anowicki on [redacted]
anowicki on [redacted]
anowicki on [redacted]
anowicki on [redacted]
anowicki on [redacted]
anowicki on [redacted]
anowicki on [redacted]
anowicki on [redacted]

anowicki on [redacted]

Login:	**anowicki**
Database:	[redacted]
Description:	[redacted]

Owner actions...
Unlock account, increase quota.

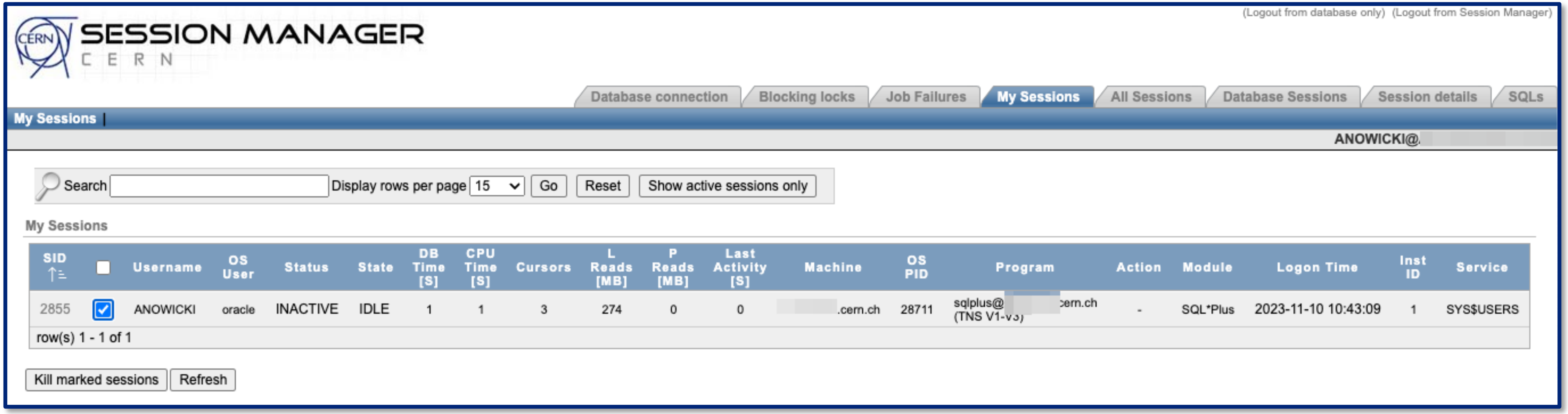
DADs...

Change Password

Change Description

Change Owner
Requires new owner's approval.

Oracle



ANOWICKI@[redacted]

My Sessions

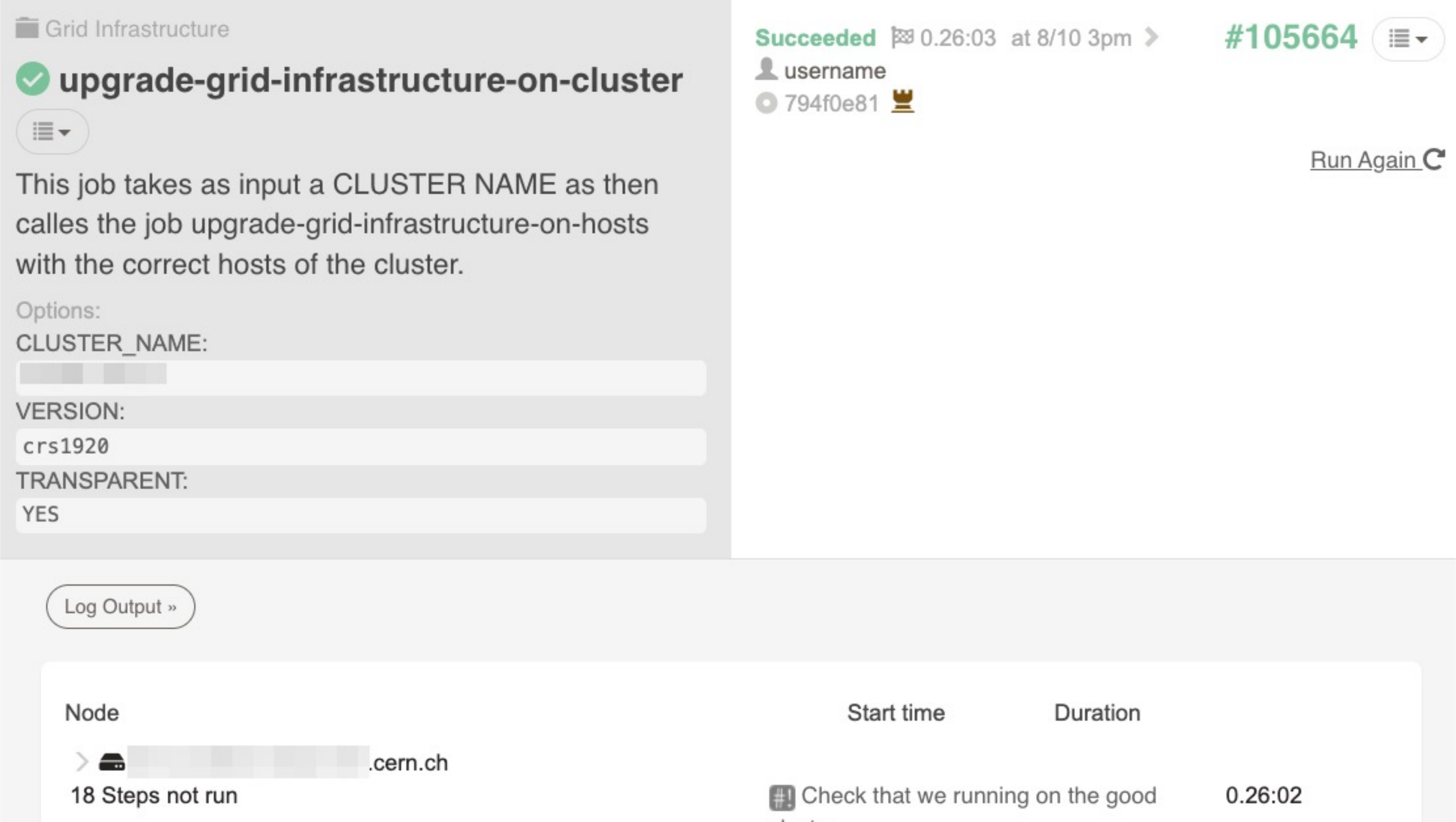
SID	☐	Username	OS User	Status	State	DB Time [S]	CPU Time [S]	Cursors	L Reads [MB]	P Reads [MB]	Last Activity [S]	Machine	OS PID	Program	Action	Module	Logon Time	Inst ID	Service
2855	☑	ANOWICKI	oracle	INACTIVE	IDLE	1	1	3	274	0	0	[redacted].cern.ch	28711	sqlplus@[redacted]cern.ch (TNS V1-V3)	-	SQL*Plus	2023-11-10 10:43:09	1	SYS$USERS

row(s) 1 - 1 of 1

Automation

Automate things that you do often:

- **Patching of databases & clusterware**
- **Hadware migrations between servers**
- **New installations**
- **Accounts provisioning (and deletion)**



Grid Infrastructure

upgrade-grid-infrastructure-on-cluster

This job takes as input a CLUSTER NAME as then calles the job upgrade-grid-infrastructure-on-hosts with the correct hosts of the cluster.

Options:

CLUSTER_NAME:
[redacted]

VERSION:
`crs1920`

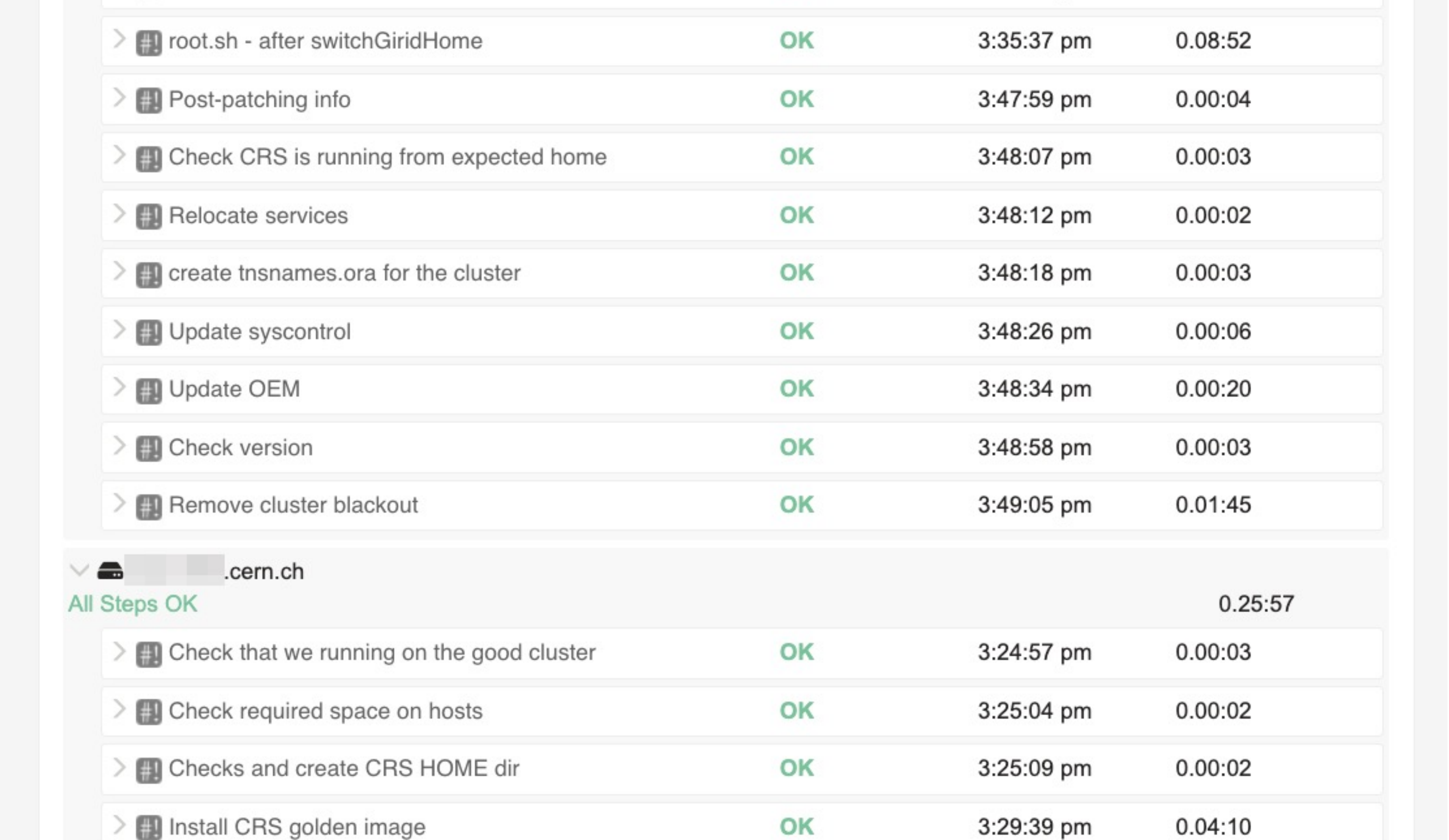
TRANSPARENT:
YES

Succeeded 0.26:03 at 8/10 3pm #105664

username

794f0e81

Node	Start time	Duration
[redacted].cern.ch		
18 Steps not run	Check that we running on the good	0.26:02

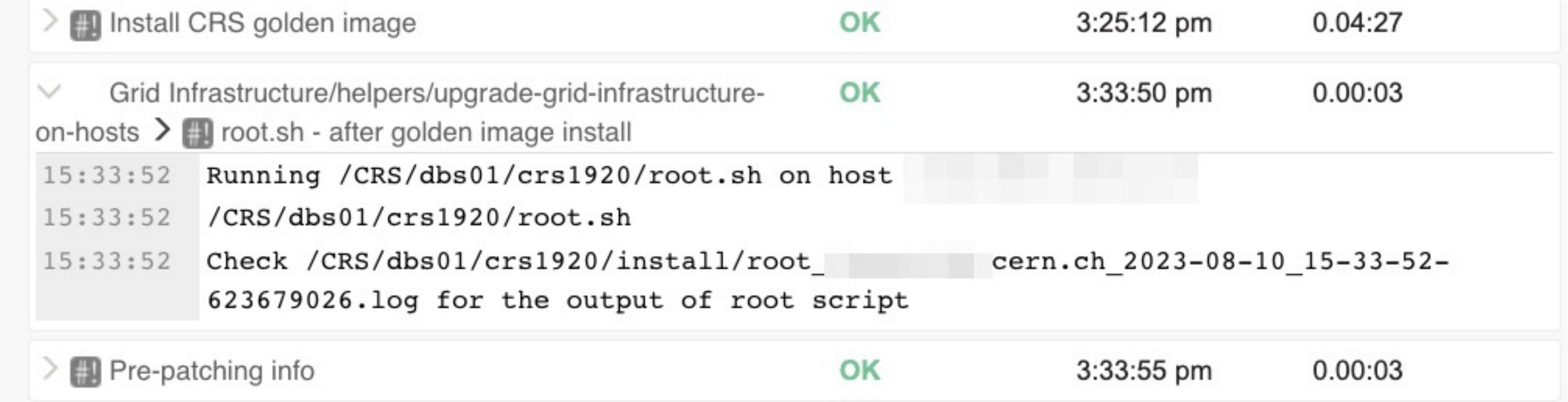


root.sh - after switchGiridHome	OK	3:35:37 pm	0.08:52
Post-patching info	OK	3:47:59 pm	0.00:04
Check CRS is running from expected home	OK	3:48:07 pm	0.00:03
Relocate services	OK	3:48:12 pm	0.00:02
create tnsnames.ora for the cluster	OK	3:48:18 pm	0.00:03
Update syscontrol	OK	3:48:26 pm	0.00:06
Update OEM	OK	3:48:34 pm	0.00:20
Check version	OK	3:48:58 pm	0.00:03
Remove cluster blackout	OK	3:49:05 pm	0.01:45

[redacted].cern.ch

All Steps OK 0.25:57

Check that we running on the good cluster	OK	3:24:57 pm	0.00:03
Check required space on hosts	OK	3:25:04 pm	0.00:02
Checks and create CRS HOME dir	OK	3:25:09 pm	0.00:02
Install CRS golden image	OK	3:29:39 pm	0.04:10



Install CRS golden image OK 3:25:12 pm 0.04:27

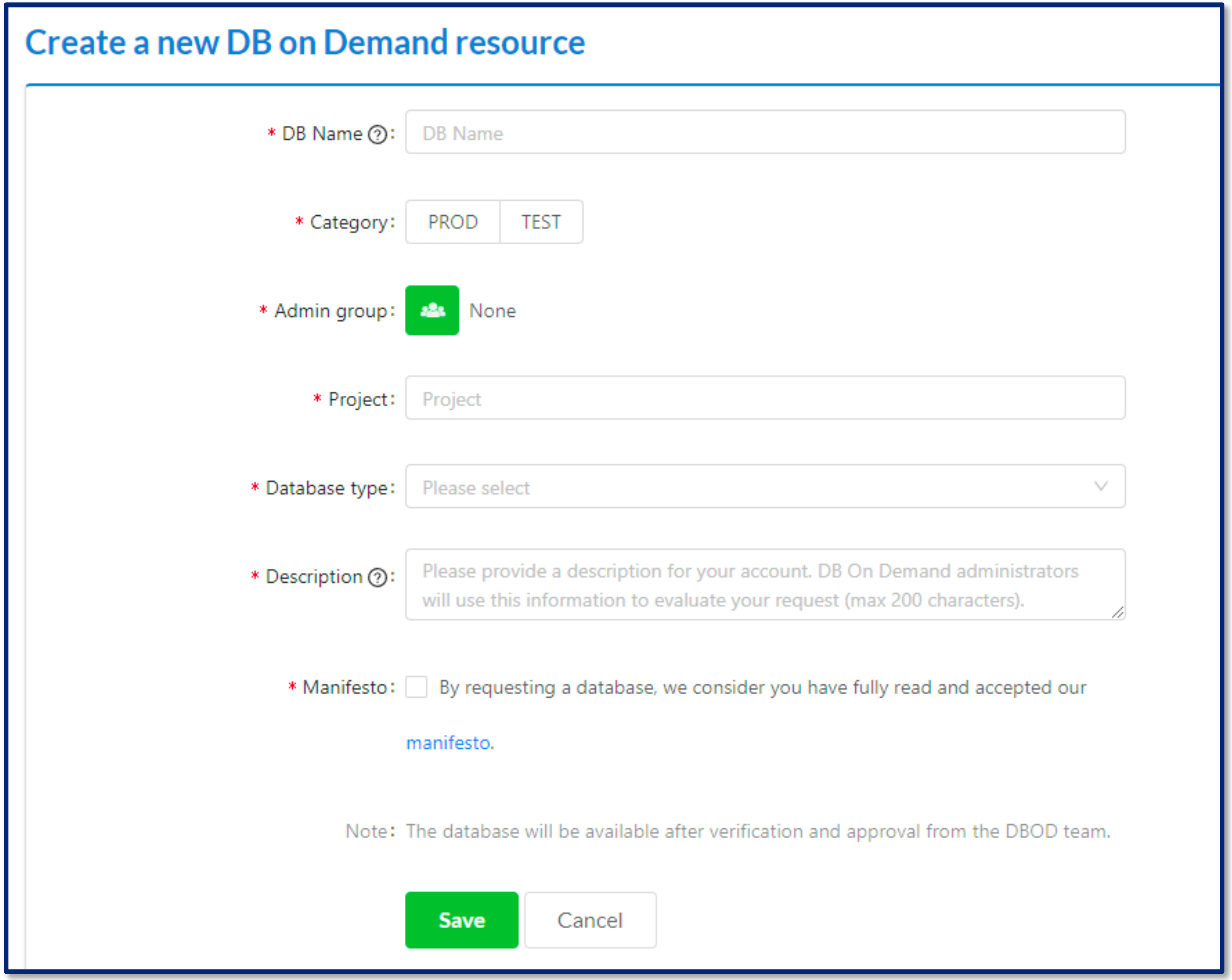
Grid Infrastructure/helpers/upgrade-grid-infrastructure-on-hosts > root.sh - after golden image install OK 3:33:50 pm 0.00:03

```
15:33:52  Running /CRS/dbs01/crs1920/root.sh on host [redacted]
15:33:52  /CRS/dbs01/crs1920/root.sh
15:33:52  Check /CRS/dbs01/crs1920/install/root_[redacted]cern.ch_2023-08-10_15-33-52-
          623679026.log for the output of root script
```

Pre-patching info OK 3:33:55 pm 0.00:03

https://github.com/rundeck/rundeck

Database on Demand



Create a new DB on Demand resource

* DB Name ⓘ: DB Name

* Category: PROD | TEST

* Admin group: None

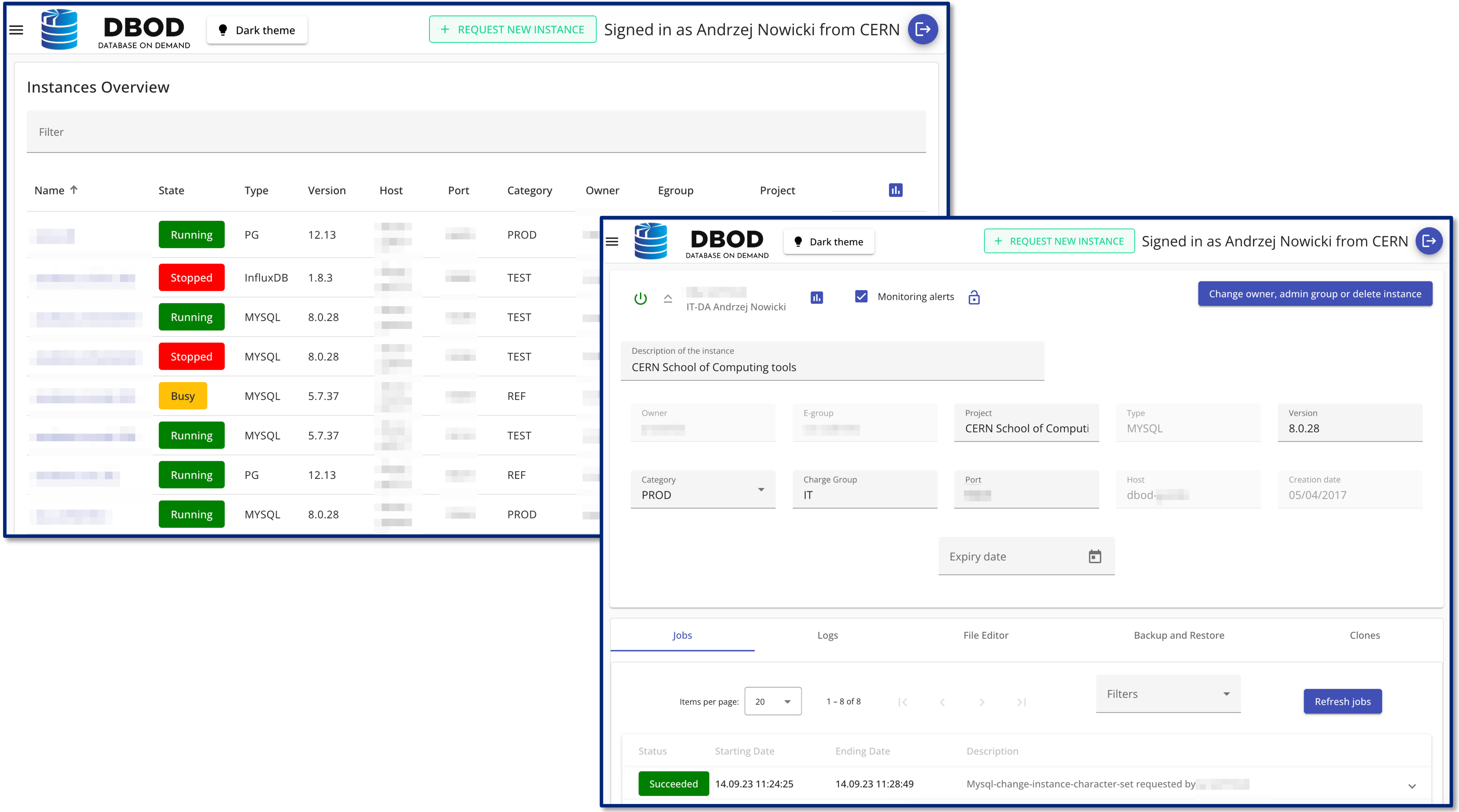
* Project: Project

* Database type: Please select

* Description ⓘ: Please provide a description for your account. DB On Demand administrators will use this information to evaluate your request (max 200 characters).

* Manifesto: □ By requesting a database, we consider you have fully read and accepted our manifesto.

Note: The database will be available after verification and approval from the DBOD team.



Instances Overview

Filter

Name ↑	State	Type	Version	Host	Port	Category	Owner	Egroup	Project
[redacted]	Running	PG	12.13	[redacted]	[redacted]	PROD			
[redacted]	Stopped	InfluxDB	1.8.3	[redacted]	[redacted]	TEST			
[redacted]	Running	MYSQL	8.0.28	[redacted]	[redacted]	TEST			
[redacted]	Stopped	MYSQL	8.0.28	[redacted]	[redacted]	TEST			
[redacted]	Busy	MYSQL	5.7.37	[redacted]	[redacted]	REF			
[redacted]	Running	MYSQL	5.7.37	[redacted]	[redacted]	TEST			
[redacted]	Running	PG	12.13	[redacted]	[redacted]	REF			
[redacted]	Running	MYSQL	8.0.28	[redacted]	[redacted]	PROD			

[redacted]
IT-DA Andrzej Nowicki

☑ Monitoring alerts

Change owner, admin group or delete instance

Description of the instance: CERN School of Computing tools

Owner: [redacted]

E-group: [redacted]

Project: CERN School of Computi

Type: MYSQL

Version: 8.0.28

Category: PROD

Charge Group: IT

Port: [redacted]

Host: dbod-[redacted]

Creation date: 05/04/2017

Expiry date

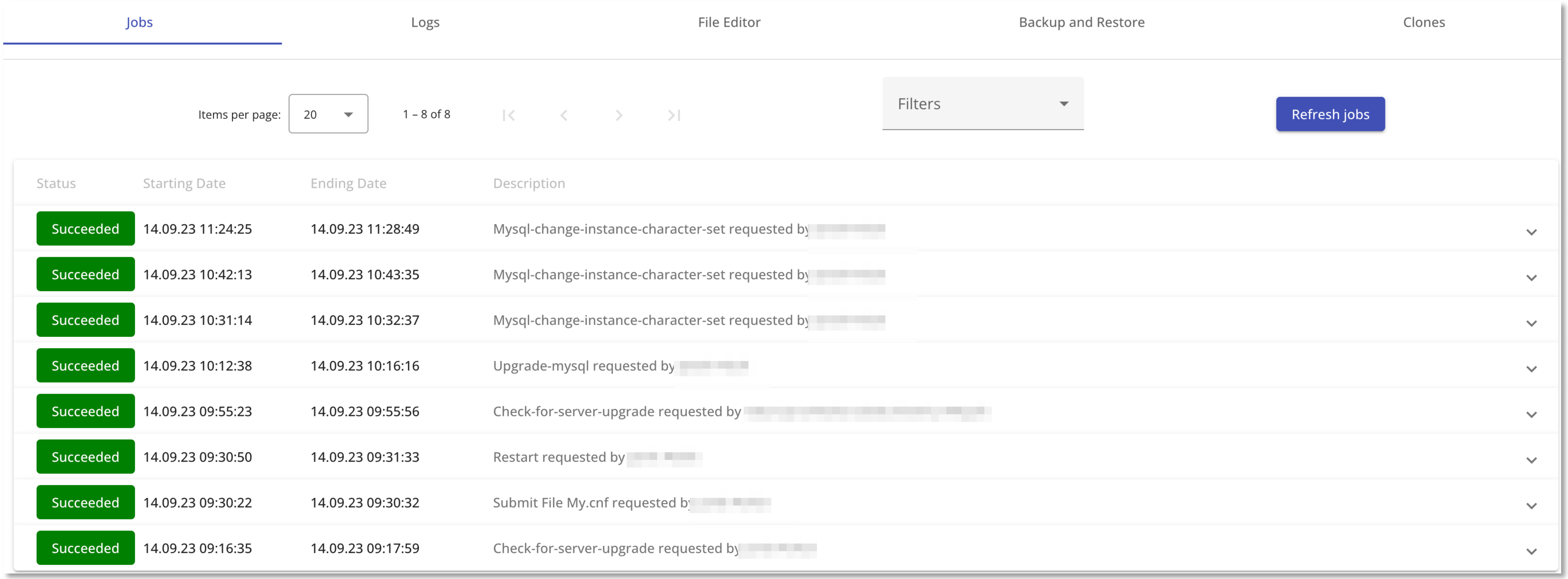
Items per page: 20

1 – 8 of 8

Filters

Refresh jobs

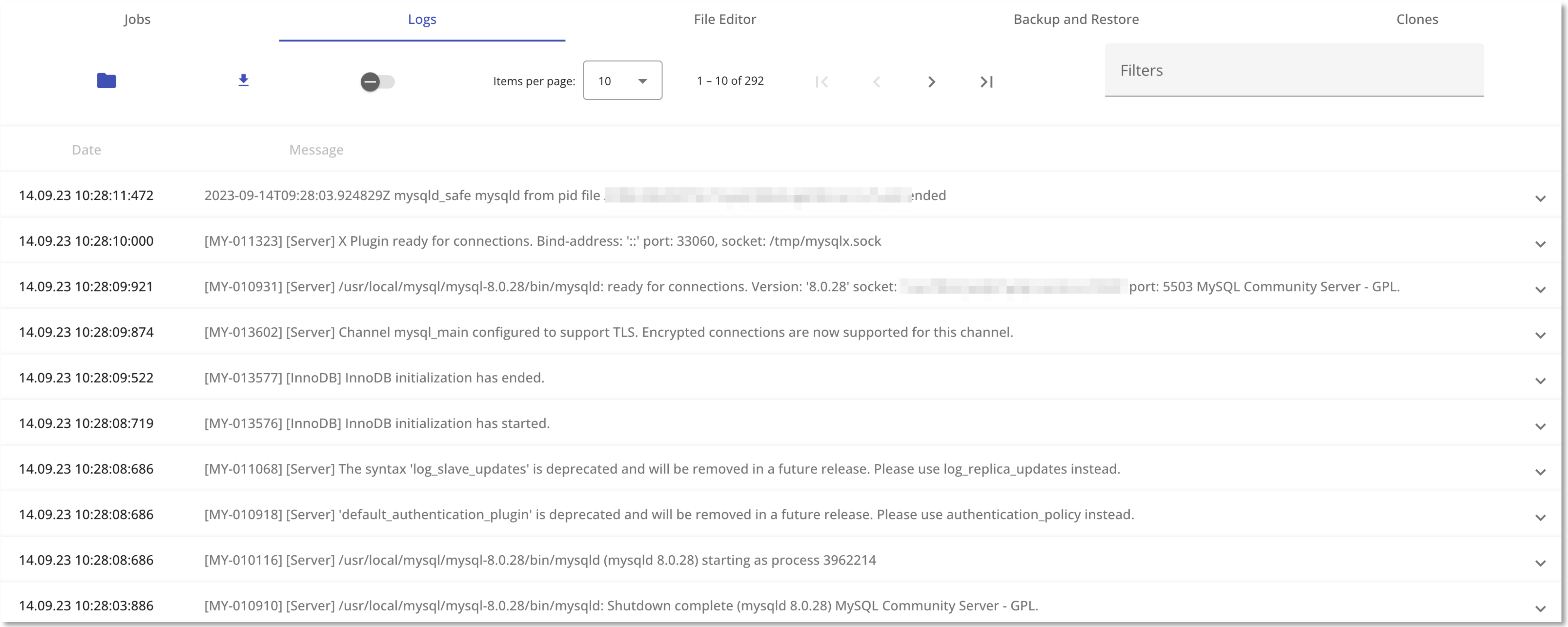
Status	Starting Date	Ending Date	Description
Succeeded	14.09.23 11:24:25	14.09.23 11:28:49	Mysql-change-instance-character-set requested by [redacted]



Items per page: 20

1 – 8 of 8

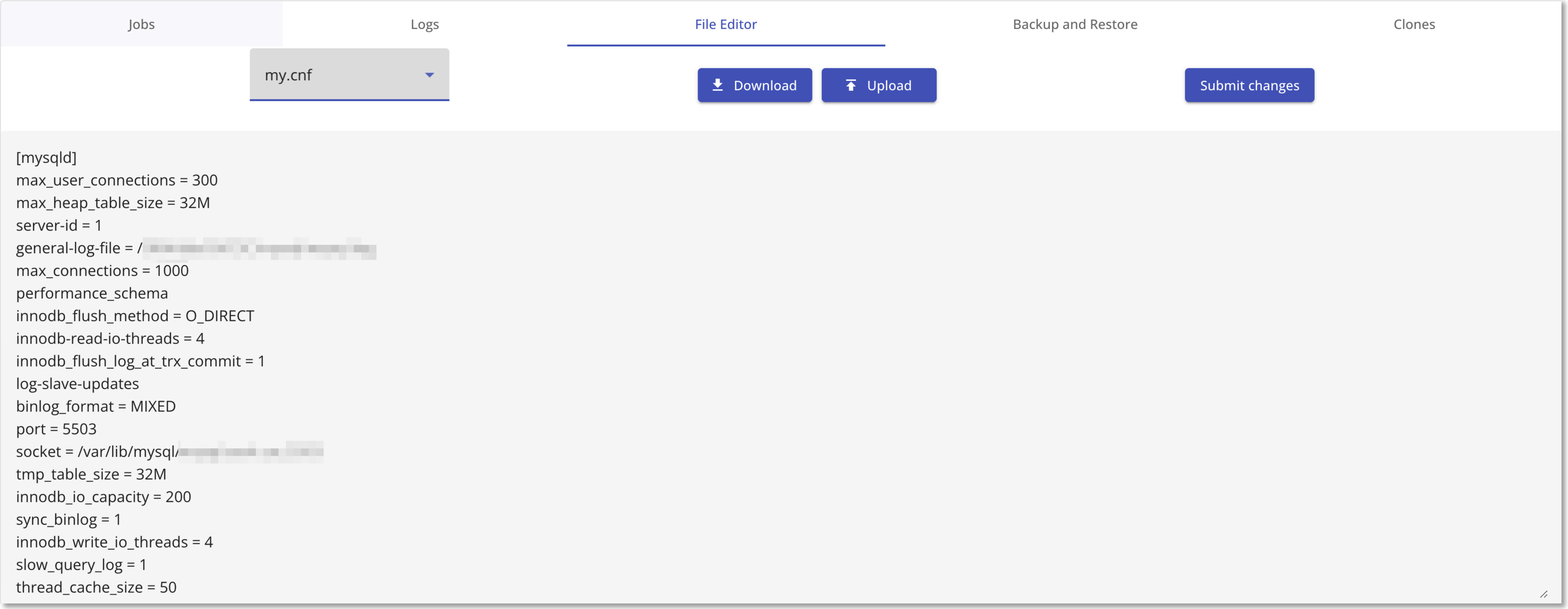
Status	Starting Date	Ending Date	Description
Succeeded	14.09.23 11:24:25	14.09.23 11:28:49	Mysql-change-instance-character-set requested by [redacted]
Succeeded	14.09.23 10:42:13	14.09.23 10:43:35	Mysql-change-instance-character-set requested by [redacted]
Succeeded	14.09.23 10:31:14	14.09.23 10:32:37	Mysql-change-instance-character-set requested by [redacted]
Succeeded	14.09.23 10:12:38	14.09.23 10:16:16	Upgrade-mysql requested by [redacted]
Succeeded	14.09.23 09:55:23	14.09.23 09:55:56	Check-for-server-upgrade requested by [redacted]
Succeeded	14.09.23 09:30:50	14.09.23 09:31:33	Restart requested by [redacted]
Succeeded	14.09.23 09:30:22	14.09.23 09:30:32	Submit File My.cnf requested by [redacted]
Succeeded	14.09.23 09:16:35	14.09.23 09:17:59	Check-for-server-upgrade requested by [redacted]



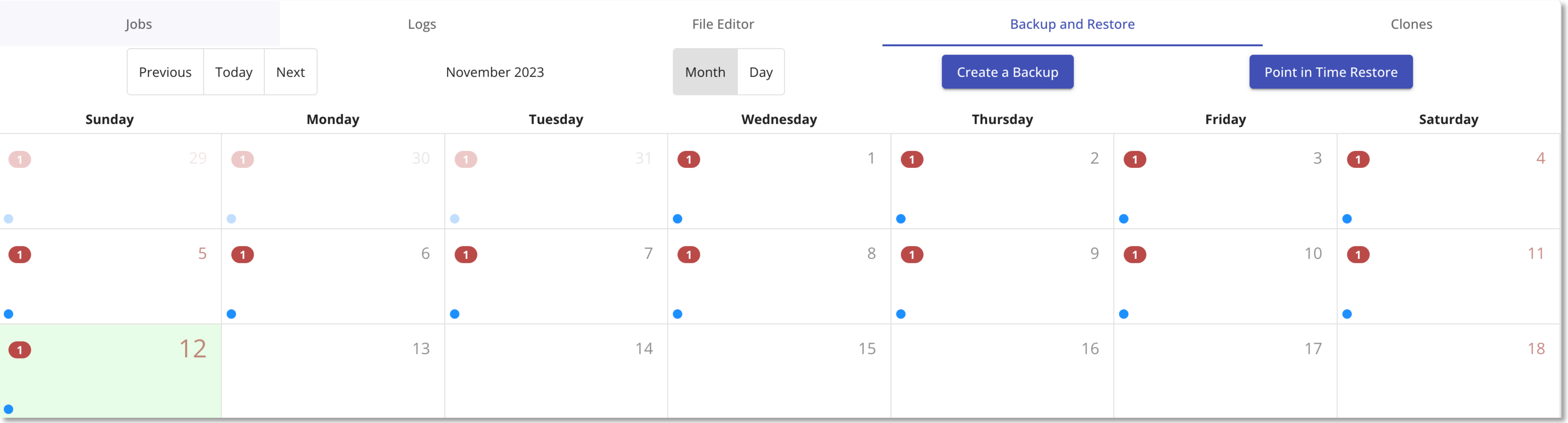
Items per page: 10

1 – 10 of 292

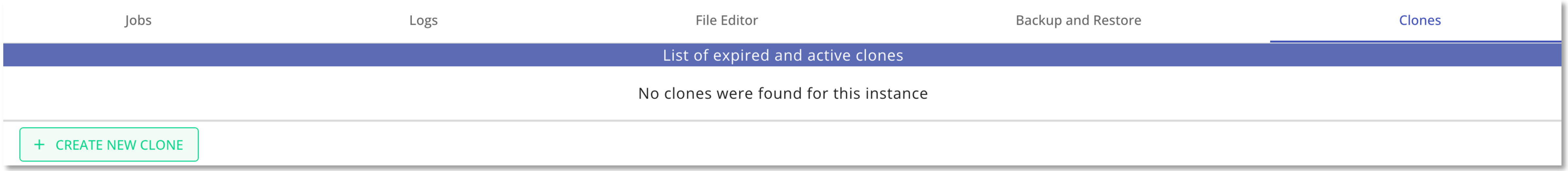
```
Date                     Message
14.09.23 10:28:11:472    2023-09-14T09:28:03.924829Z mysqld_safe mysqld from pid file [redacted]nded
14.09.23 10:28:10:000    [MY-011323] [Server] X Plugin ready for connections. Bind-address: '::' port: 33060, socket: /tmp/mysqlx.sock
14.09.23 10:28:09:921    [MY-010931] [Server] /usr/local/mysql/mysql-8.0.28/bin/mysqld: ready for connections. Version: '8.0.28' socket: [redacted] port: 5503 MySQL Community Server - GPL.
14.09.23 10:28:09:874    [MY-013602] [Server] Channel mysql_main configured to support TLS. Encrypted connections are now supported for this channel.
14.09.23 10:28:09:522    [MY-013577] [InnoDB] InnoDB initialization has ended.
14.09.23 10:28:08:719    [MY-013576] [InnoDB] InnoDB initialization has started.
14.09.23 10:28:08:686    [MY-011068] [Server] The syntax 'log_slave_updates' is deprecated and will be removed in a future release. Please use log_replica_updates instead.
14.09.23 10:28:08:686    [MY-010918] [Server] 'default_authentication_plugin' is deprecated and will be removed in a future release. Please use authentication_policy instead.
14.09.23 10:28:08:686    [MY-010116] [Server] /usr/local/mysql/mysql-8.0.28/bin/mysqld (mysqld 8.0.28) starting as process 3962214
14.09.23 10:28:03:886    [MY-010910] [Server] /usr/local/mysql/mysql-8.0.28/bin/mysqld: Shutdown complete (mysqld 8.0.28) MySQL Community Server - GPL.
```



```
[mysqld]
max_user_connections = 300
max_heap_table_size = 32M
server-id = 1
general-log-file = /[redacted]
max_connections = 1000
performance_schema
innodb_flush_method = O_DIRECT
innodb-read-io-threads = 4
innodb_flush_log_at_trx_commit = 1
log-slave-updates
binlog_format = MIXED
port = 5503
socket = /var/lib/mysql/[redacted]
tmp_table_size = 32M
innodb_io_capacity = 200
sync_binlog = 1
innodb_write_io_threads = 4
slow_query_log = 1
thread_cache_size = 50
```

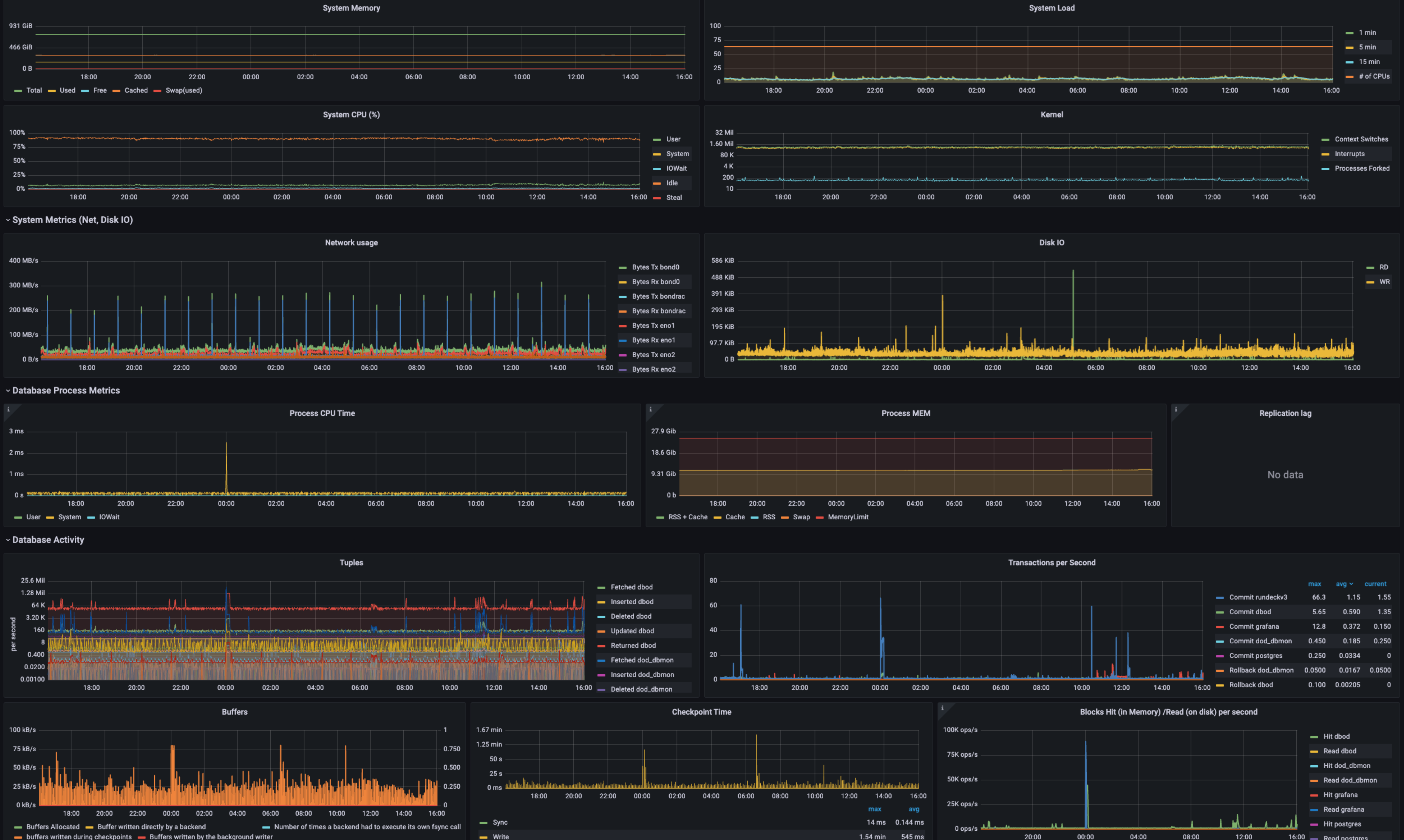


Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1 29 ●	1 30 ●	1 31 ●	1 1 ●	1 2 ●	1 3 ●	1 4 ●
1 5 ●	1 6 ●	1 7 ●	1 8 ●	1 9 ●	1 10 ●	1 11 ●
1 12 ●	13	14	15	16	17	18



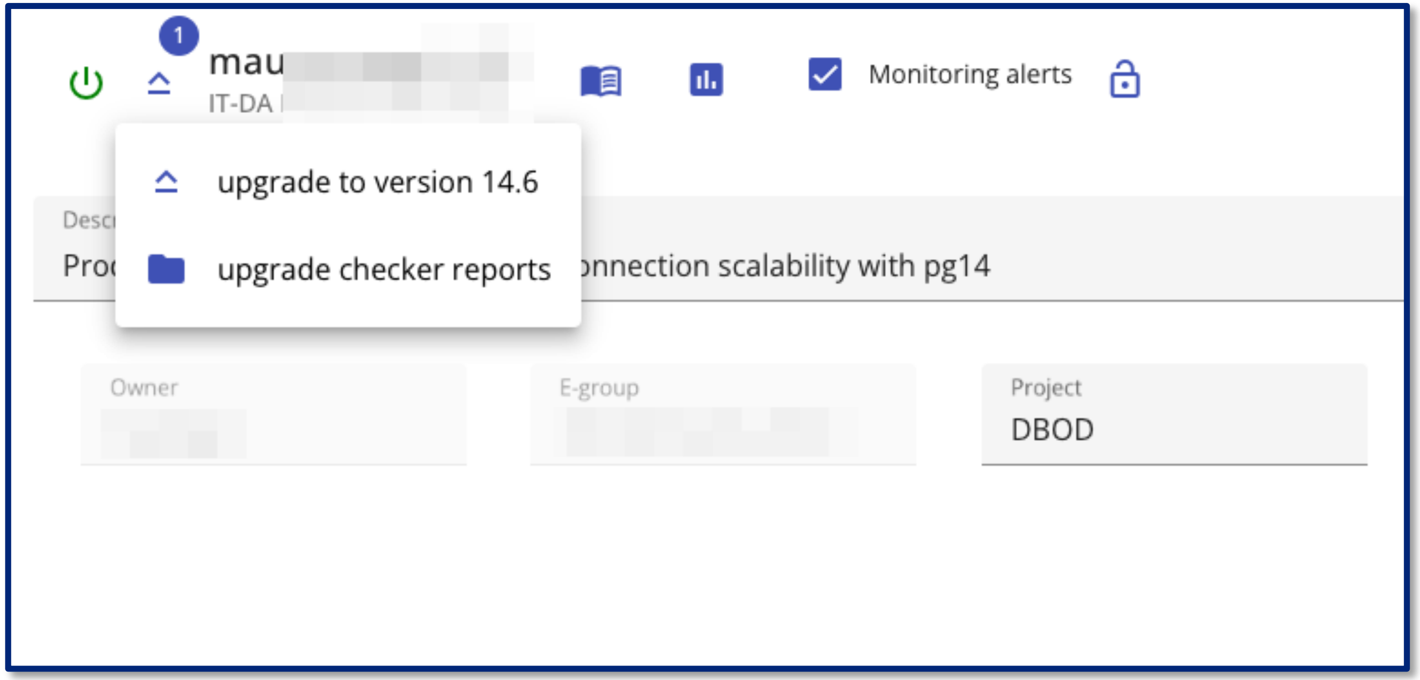
List of expired and active clones

No clones were found for this instance



	max	avg	current
Commit rundeckv3	66.3	1.15	1.55
Commit dbod	5.65	0.590	1.35
Commit grafana	12.8	0.372	0.150
Commit dod_dbmon	0.450	0.185	0.250
Commit postgres	0.250	0.0334	0
Rollback dod_dbmon	0.0500	0.0167	0.0500
Rollback dbod	0.100	0.00205	0

	max	avg
Sync	14 ms	0.144 ms
Write	1.54 min	545 ms



mau[redacted]

IT-DA [redacted]

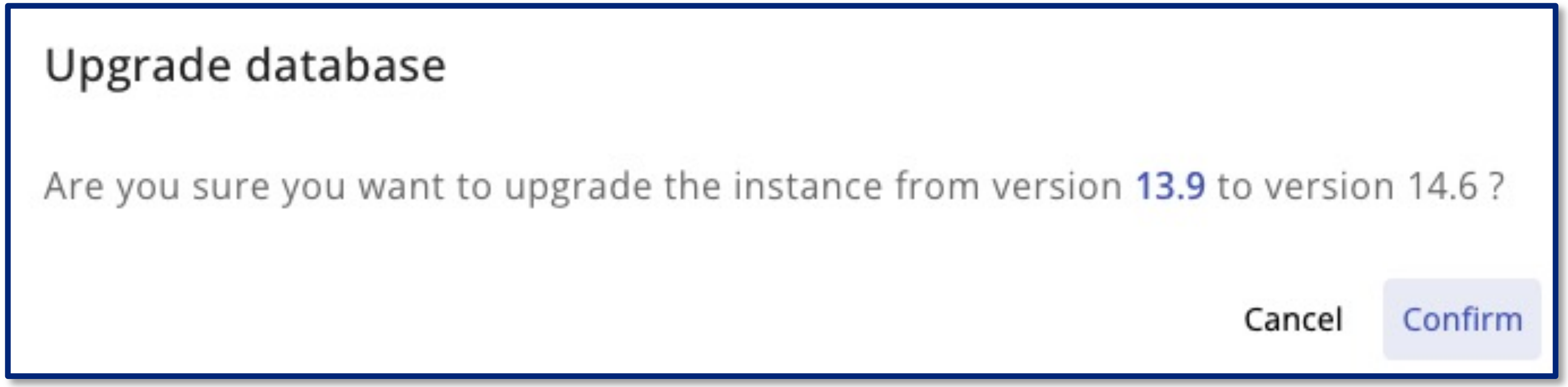
Desc

Proc onnection scalability with pg14

Owner [redacted]

E-group [redacted]

Project DBOD



Upgrade database

Are you sure you want to upgrade the instance from version 13.9 to version 14.6 ?

Services relying on DBoD

- Indico (meeting planning, room booking system)
- Autorization Service (SSO, 2FA)
- Configuration Management (puppet, foreman), Secrets Vault
- Jira, Gitlab, Openstack
- Websites (https://home.cern/)
- CERN Document Server (https://cds.cern.ch)
- File Transfer System
- ATLAS Panda (Production and Distributed Analysis System)
- and more…

UPGRADE
ALL THE THINGS
memegenerator.net

PostgreSQL is sensitive to glibc locale changes

RHEL/CentOS 9 introduces new glibc, with new locale

Sorting order is different with new locale

Postgres uses glibc sorting to create indexes

This might lead to broken indexes

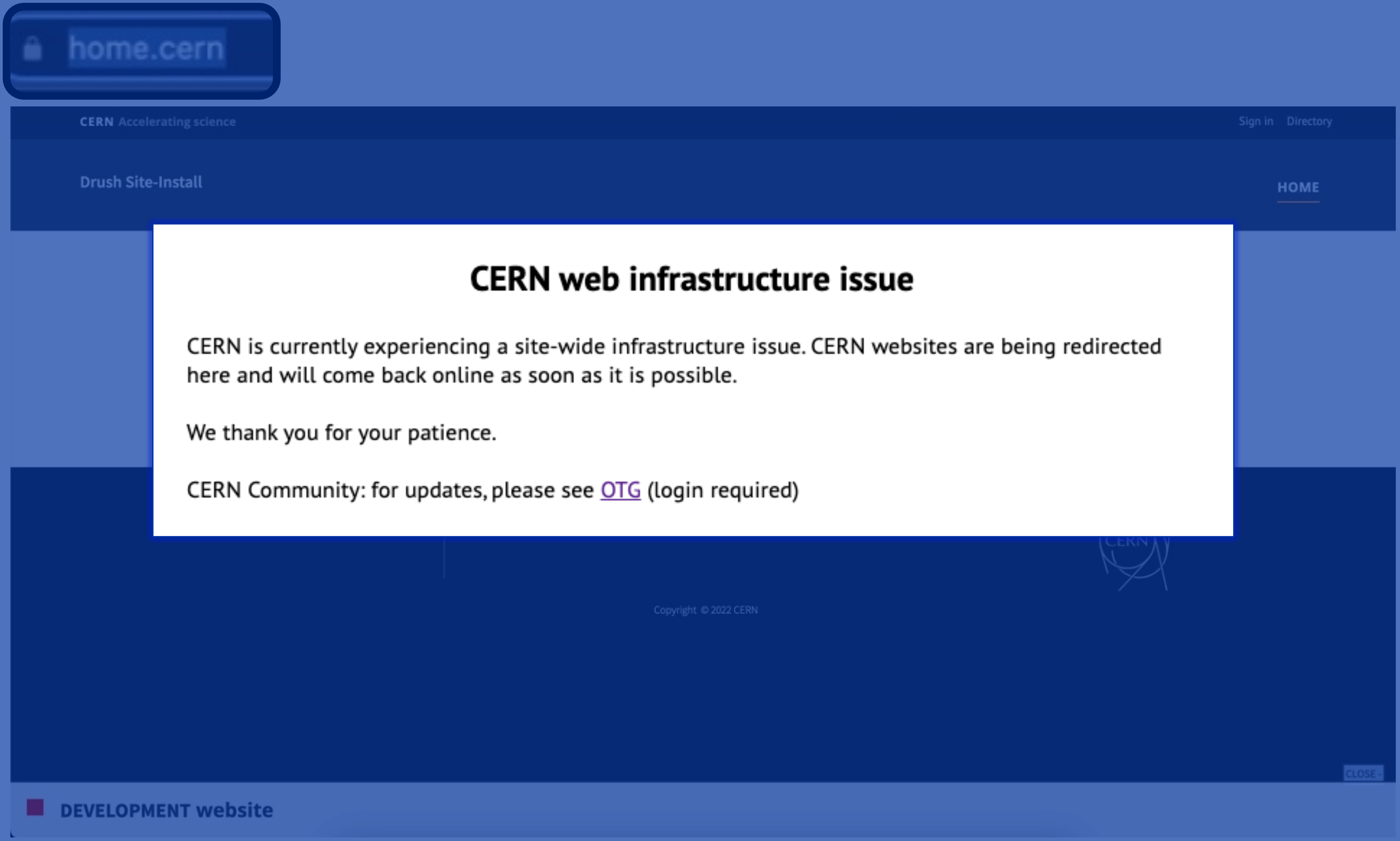
We detect it and automate the rebuild of affected indexes for our users

https://wiki.postgresql.org/wiki/Locale_data_changes
https://postgresql.verite.pro/blog/2018/08/27/glibc-upgrade.html

Story time!

second to last working day in 2022

20 December 2022 15:00



CERN web infrastructure issue

CERN is currently experiencing a site-wide infrastructure issue. CERN websites are being redirected here and will come back online as soon as it is possible.

We thank you for your patience.

CERN Community: for updates, please see OTG (login required)

Investigation

- **The webservers are in a weird state, serving empty pages**
- **Some databases are not there anymore**

What happened?

- **Orchestration deployed QA settings on PROD**
- **This lead to removal of some dependencies**
- **Which lead to cleanup of "*unused*" resources using an API we provide**

Oops!

Automate as much as possible – caveats

- **Protect critical infrastructure from automatic deletion**
- **Implement some combination of:**
  - rate limiting
  - having an approval process for the removal of production resources
  - disabling the auto removal of mission critical resources
- **Expect e.g:**
  - resource management returning empty data 😅
  - orchestration removing configuration and linked resources, such as… **databases**

There’s more

Corner cases

Consider corner cases. What will happen if you try to upgrade:

- resource management database?
- automation tool (e.g. rundeck) internal database?
- configuration management tool internal database (e.g. puppet)?
- database of the authorization service?

Will your automation still work?

Corner cases

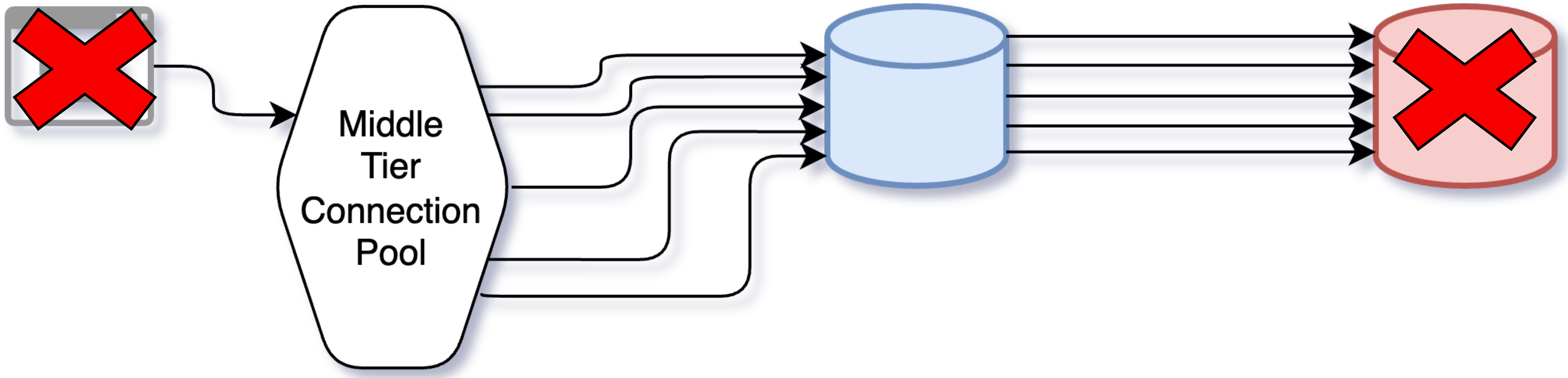
Do you use DNS names to configure database replication?

- What if your DNS service is relying on your databases?

- If you do, make sure you have entries in `/etc/hosts`

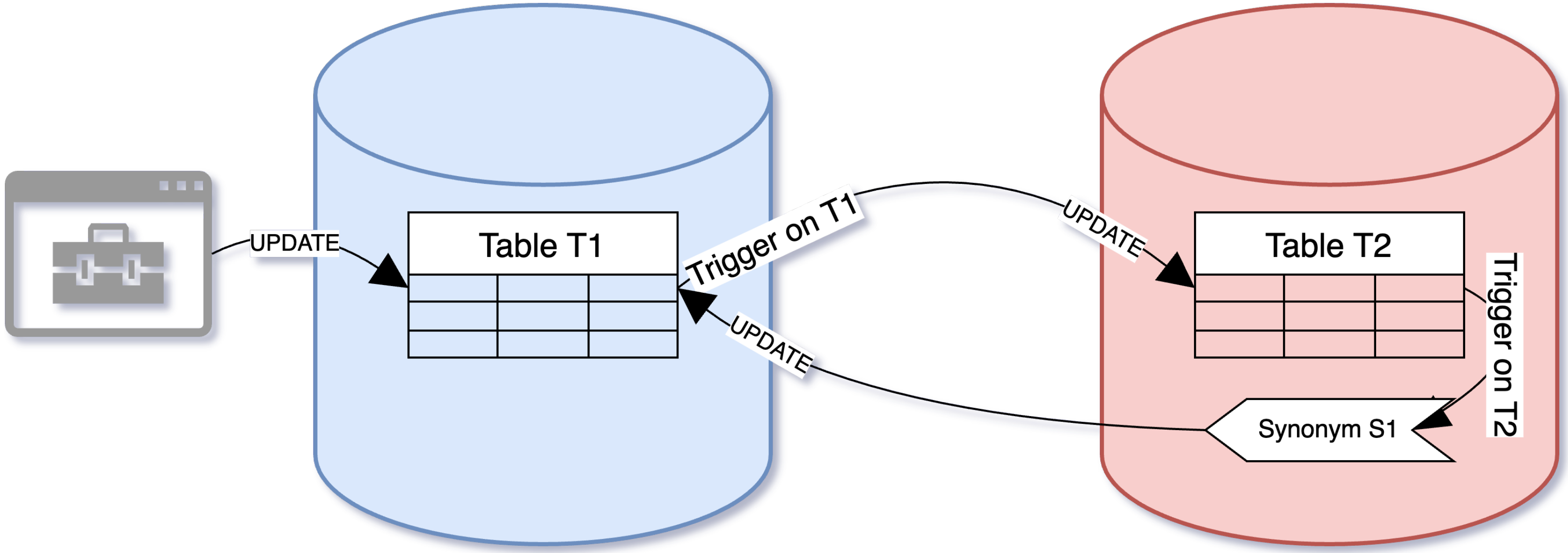
Database Links

What if you shutdown one of the db’s involved?



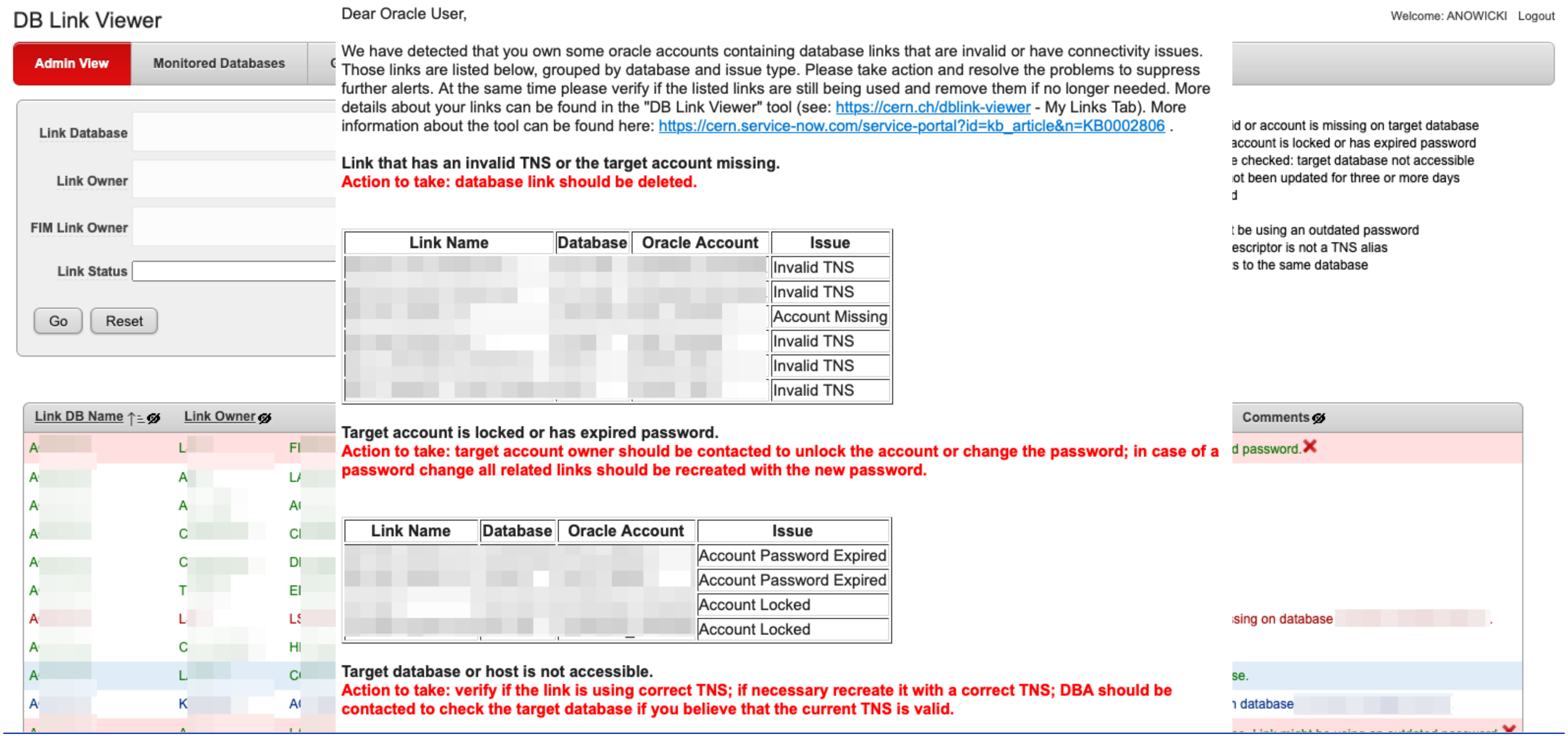
Database Links

Triggers on remote tables having triggers on remote tables?



Database Links

Database Links


DB Link Viewer
Welcome: ANOWICKI Logout
Admin View
Monitored Databases
Link Database
Link Owner
FIM Link Owner
Link Status
Go
Reset
id or account is missing on target database
account is locked or has expired password
e checked: target database not accessible
ot been updated for three or more days
d
t be using an outdated password
escriptor is not a TNS alias
s to the same database
Link DB Name
Link Owner
Comments
d password.
sing on database
se.
n database

Dear Oracle User,

We have detected that you own some oracle accounts containing database links that are invalid or have connectivity issues. Those links are listed below, grouped by database and issue type. Please take action and resolve the problems to suppress further alerts. At the same time please verify if the listed links are still being used and remove them if no longer needed. More details about your links can be found in the "DB Link Viewer" tool (see: https://cern.ch/dblink-viewer - My Links Tab). More information about the tool can be found here: https://cern.service-now.com/service-portal?id=kb_article&n=KB0002806 .

Link that has an invalid TNS or the target account missing.
Action to take: database link should be deleted.

Link Name	Database	Oracle Account	Issue
[redacted]	[redacted]	[redacted]	Invalid TNS
[redacted]	[redacted]	[redacted]	Invalid TNS
[redacted]	[redacted]	[redacted]	Account Missing
[redacted]	[redacted]	[redacted]	Invalid TNS
[redacted]	[redacted]	[redacted]	Invalid TNS
[redacted]	[redacted]	[redacted]	Invalid TNS

Target account is locked or has expired password.
Action to take: target account owner should be contacted to unlock the account or change the password; in case of a password change all related links should be recreated with the new password.

Link Name	Database	Oracle Account	Issue
[redacted]	[redacted]	[redacted]	Account Password Expired
[redacted]	[redacted]	[redacted]	Account Password Expired
[redacted]	[redacted]	[redacted]	Account Locked
[redacted]	[redacted]	[redacted]	Account Locked

Target database or host is not accessible.
Action to take: verify if the link is using correct TNS; if necessary recreate it with a correct TNS; DBA should be contacted to check the target database if you believe that the current TNS is valid.

Materialized views

```
SQL> create table tab1 (a number);

Table created.

SQL> create view v1 as select * from tab1;

View created.

SQL> create materialized view mv1 as select * from v1;

Materialized view created.

SQL> create or replace view v1 as select * from tab1 union all select * from mv1;

View created.
```

Materialized views – circular dependencies?

```
SQL> alter table tab1 modify (a number(10,1));
```

This invalidates the view

```
Table altered.

SQL> EXEC UTL_RECOMP.recomp_serial(); @?/rdbms/admin/utlrp.sql

*
ERROR at line 1:
ORA-32044: cycle detected while executing recursive WITH query
ORA-06512: at "SYS.UTL_RECOMP", line 927
ORA-06512: at "SYS.UTL_RECOMP", line 537
ORA-06512: at "SYS.UTL_RECOMP", line 896
ORA-06512: at "SYS.UTL_RECOMP", line 940
ORA-06512: at line 1
```

Materialized views – circular dependencies?

Materialized views – circular dependencies?

ORA-32044 error during the execution of ULTR.SQL (Doc ID 2592821.1)

```
SQL> -- query from 2592821.1

OWNER      OBJECT_NAME      OBJECT_ID OBJECT_TYPE.         STATUS
---------- ----------- -------------- -------------------- -------
ANOWICKI   MV1                 330520 MATERIALIZED VIEW    INVALID
```

Patching

DOs:

- **Automate as much as possible**
- **Run ”fresh” versions**
- **Look for recommendations from vendors, user communities, etc.:**
  - Oracle Database 19c Important Recommended One-off Patches (Doc ID 555.1)
  - Data Pump Recommended Proactive Patches For 19.10 and Above (Doc ID 2819284.1)
  - Latest GoldenGate/Database (OGG/RDBMS) Patch recommendations (Doc ID 2193391.1)
  - Oracle Text Mandatory and Recommended Patches (Doc ID 2644957.1)
  - https://mikedietrichde.com/

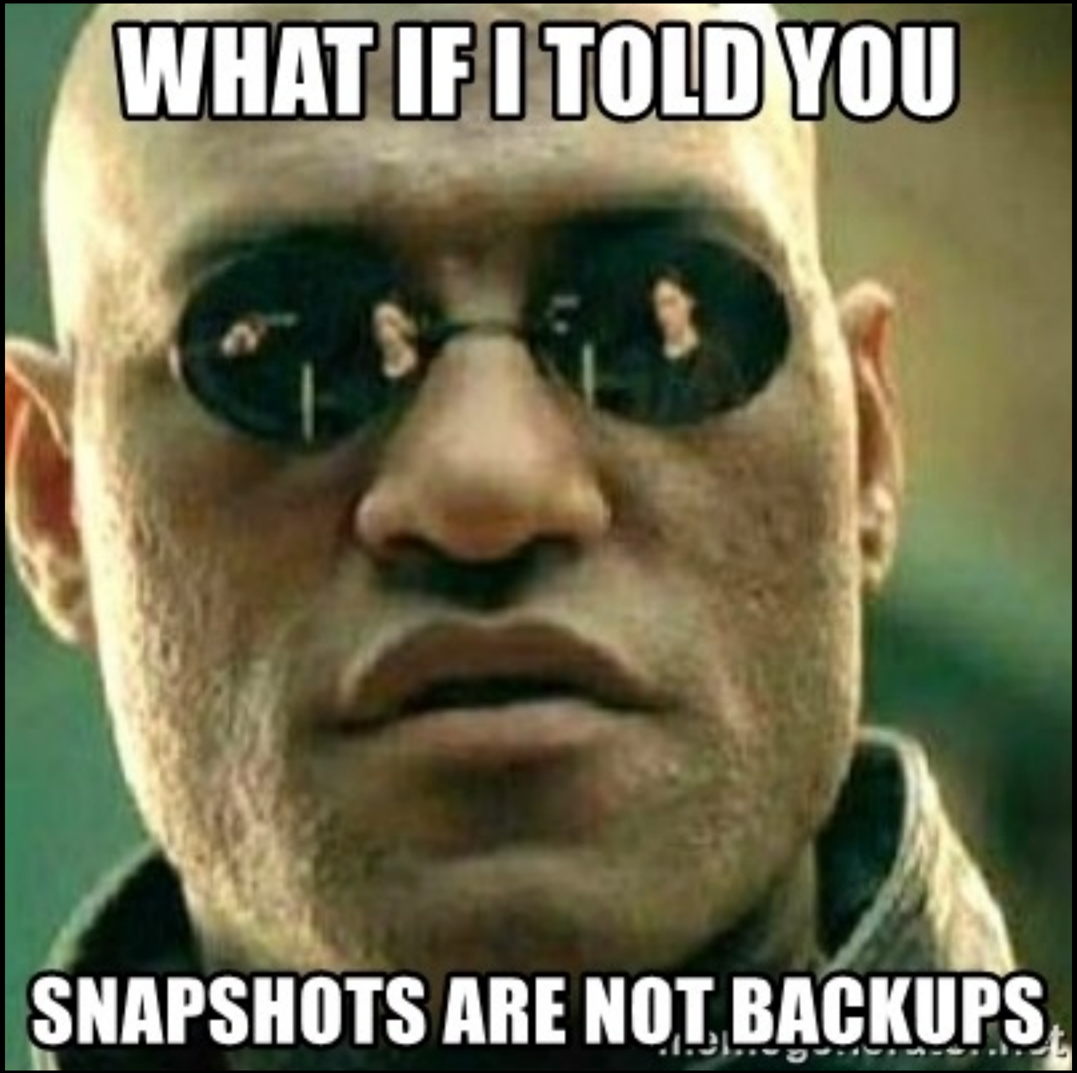
Backup

DOs:

- Use a combination of snapshot & backup
- Automatic recoveries or recovery environment
- Consider immutable backups

DON’Ts:

- Snapshots ARE NOT backup


WHAT IF I TOLD YOU
SNAPSHOTS ARE NOT BACKUPS
memegenerator.net

Very Large DBs

DOs:

- **Use partitioning: range, list, manual, interval**

Oracle VLDB and Partitioning Guide

Data Migrations

Oracle provides a tool Data Pump which speeds up copying of data by using server-side processes and internal optimisations

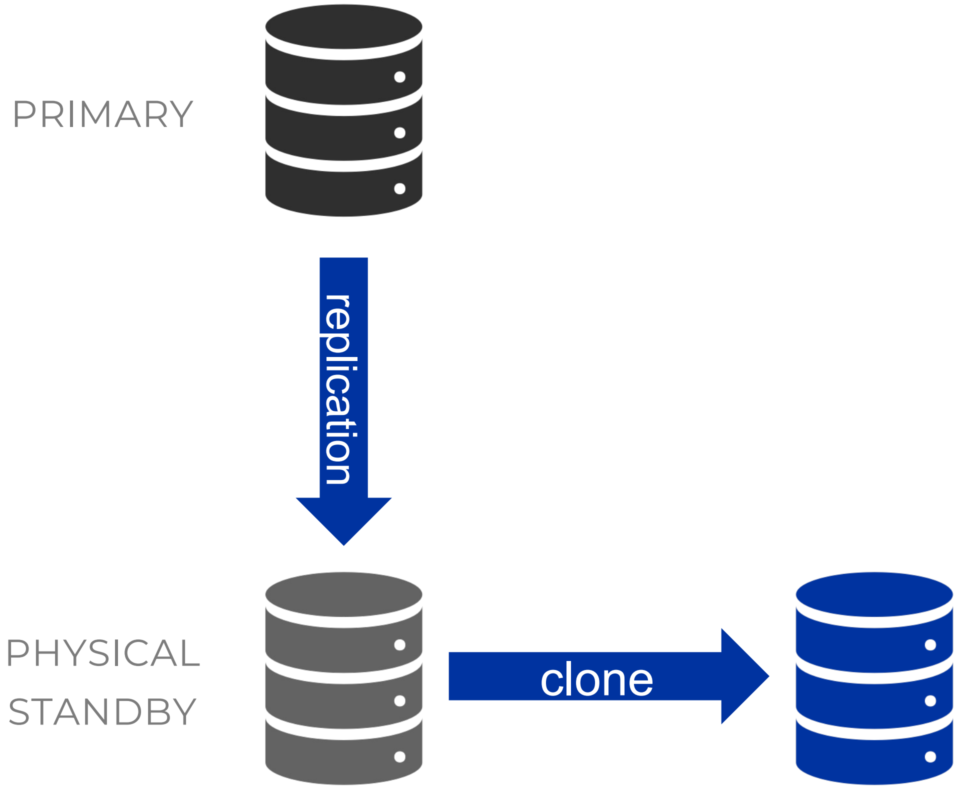
Data Migrations – Data Pump automation

```
SQL> exec cerndb_dpuser.cp_schema('DST_DB_NAME');

SQL> exec cerndb_dpuser.cp_schema('DST_DB_NAME',
               include=>'TABLE: like ''TEST%''',
               exclude=>'TABLE/CONSTRAINT;STATISTICS');
```

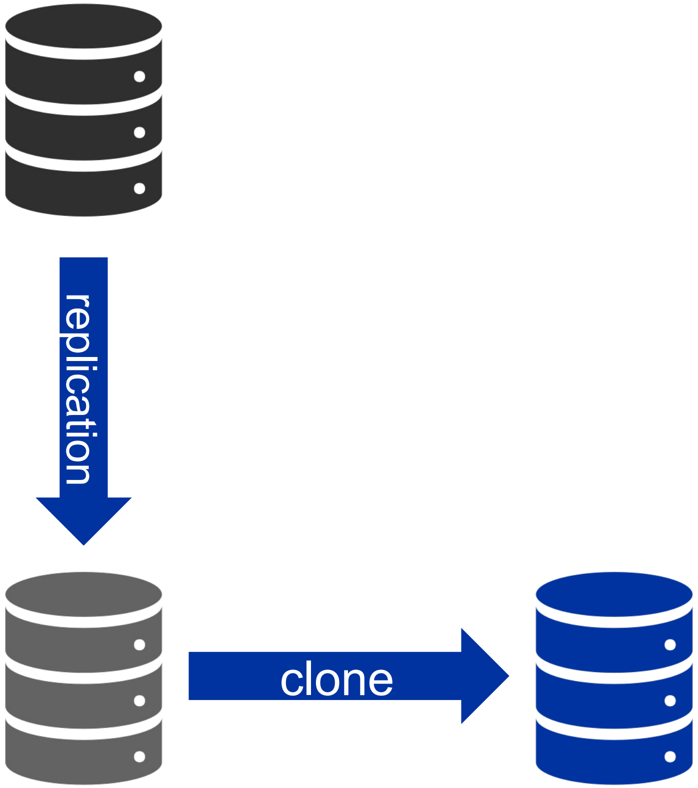
Database Clones

Using Standby DBs as a source of consistent datafiles for Clones



Database Clones

Key points



- 4 minutes to create a clone of a 10TB database
- Thin Clones (Copy On Write)

  using the dNFS snapshots

Database Clones

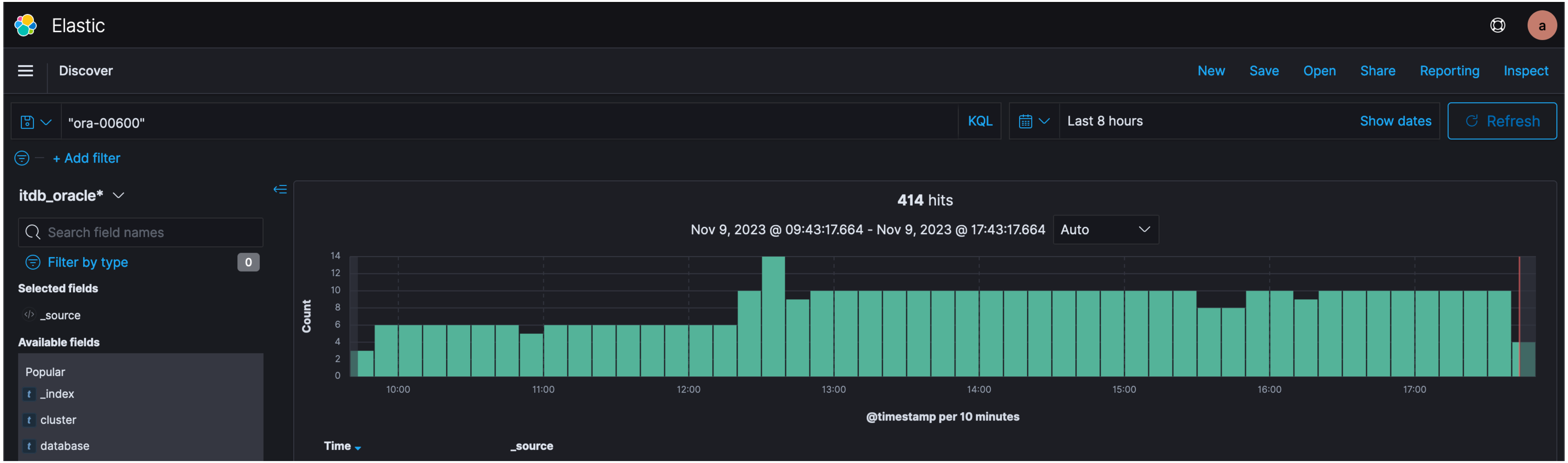
There are other possibilities for Oracle:

- Oracle Multitenant Pluggable Database Snapshot Cloning: Use Cases and Supported Platforms
  (Doc ID 1597027.1)

- Example for Cloning PDB from NON-CDB via dblink
  (Doc ID 1928653.1)

Centralised logging

Collect backup logs, audit logs & alert logs (use the xml version to filter by message level)



"ora-00600"

Last 8 hours

itdb_oracle*

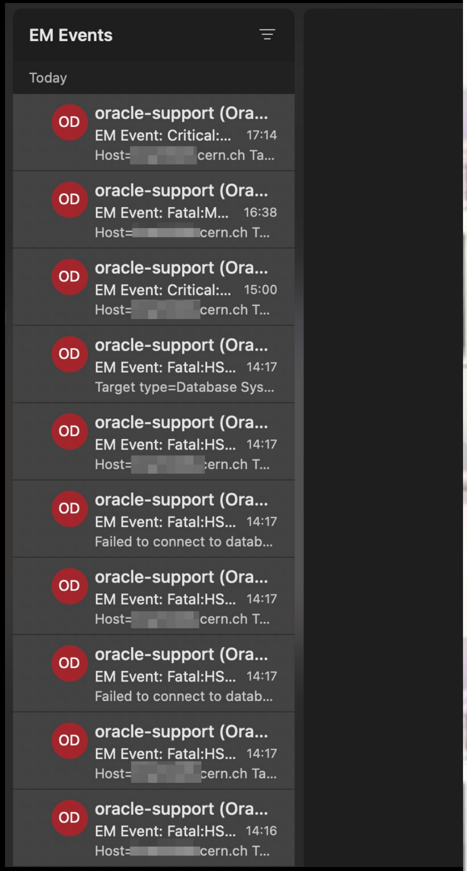
414 hits

Nov 9, 2023 @ 09:43:17.664 - Nov 9, 2023 @ 17:43:17.664

Auto

DON’T do alerting via e-mail

30 days



EM Events

Today

OD oracle-support (Ora... EM Event: Critical:... 17:14 Host=[redacted]cern.ch Ta...

OD oracle-support (Ora... EM Event: Fatal:M... 16:38 Host=[redacted]cern.ch T...

OD oracle-support (Ora... EM Event: Critical:... 15:00 Host=[redacted]cern.ch T...

OD oracle-support (Ora... EM Event: Fatal:HS... 14:17 Target type=Database Sys...

OD oracle-support (Ora... EM Event: Fatal:HS... 14:17 Host=[redacted]ern.ch T...

OD oracle-support (Ora... EM Event: Fatal:HS... 14:17 Failed to connect to datab...

OD oracle-support (Ora... EM Event: Fatal:HS... 14:17 Host=[redacted]cern.ch T...

OD oracle-support (Ora... EM Event: Fatal:HS... 14:17 Failed to connect to datab...

OD oracle-support (Ora... EM Event: Fatal:HS... 14:17 Host=[redacted]cern.ch Ta...

OD oracle-support (Ora... EM Event: Fatal:HS... 14:16 Host=[redacted]cern.ch T...

Thank you !

CERN

andrzejnowicki

andrzej.nowicki@cern.ch

www.andrzejnowicki.pl