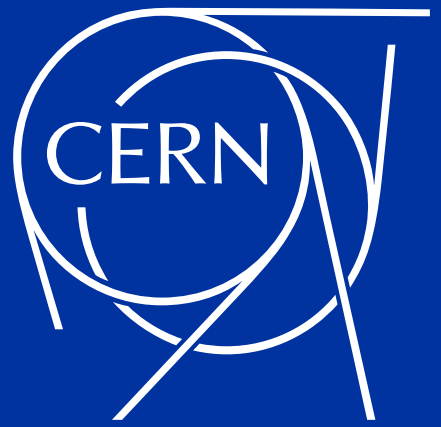


DOs and DON'Ts of managing numerous very large databases

A summary of challenges and problems encountered at CERN

Andrzej Nowicki

UKOUG 2023

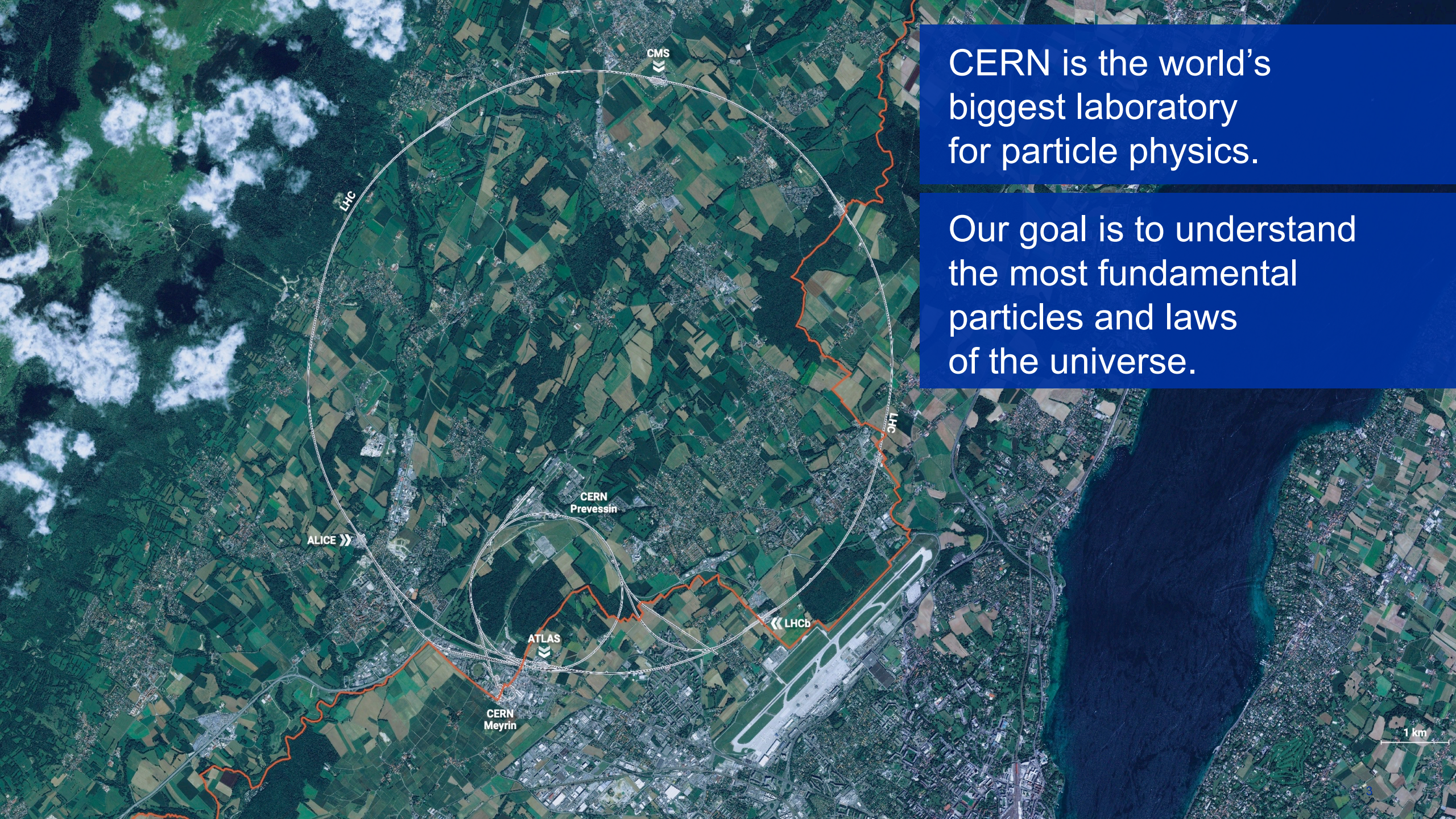
Andrzej Nowicki

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

andrzejnowicki

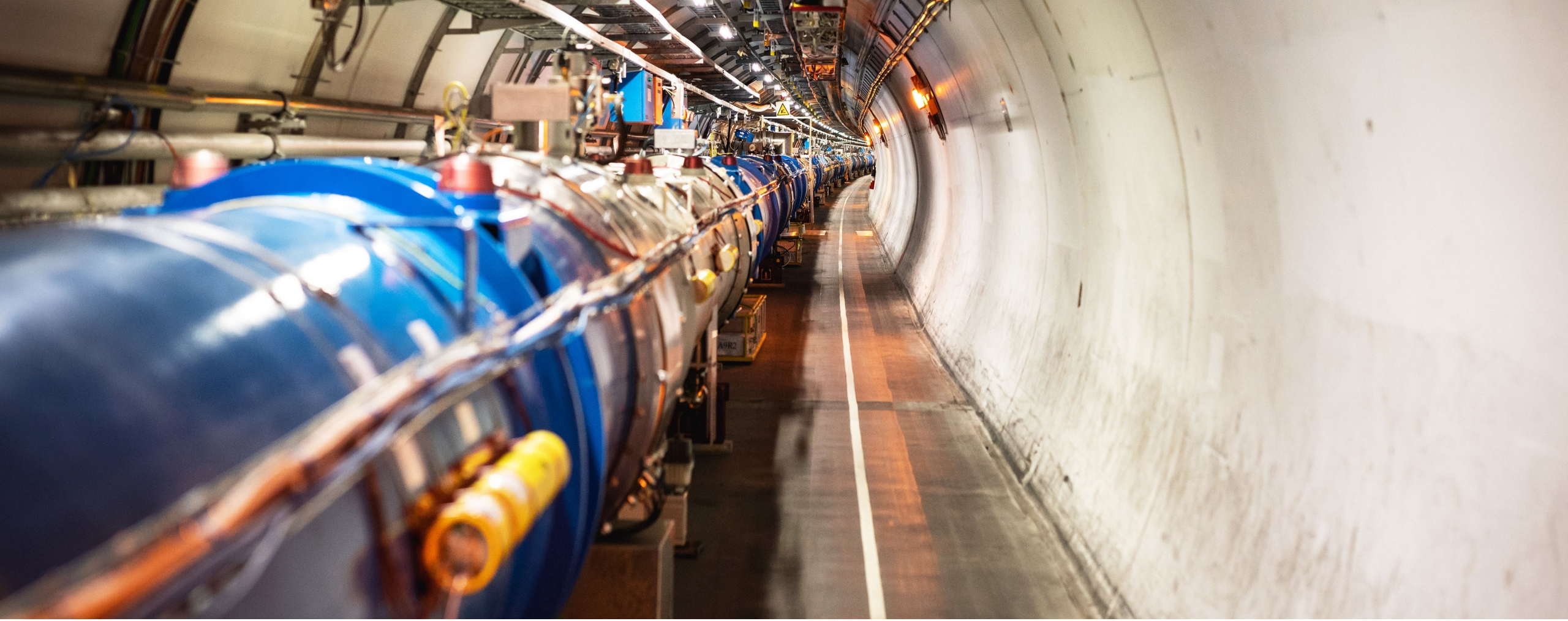
andrzej.nowicki@cern.ch

www.andrzejnowicki.pl

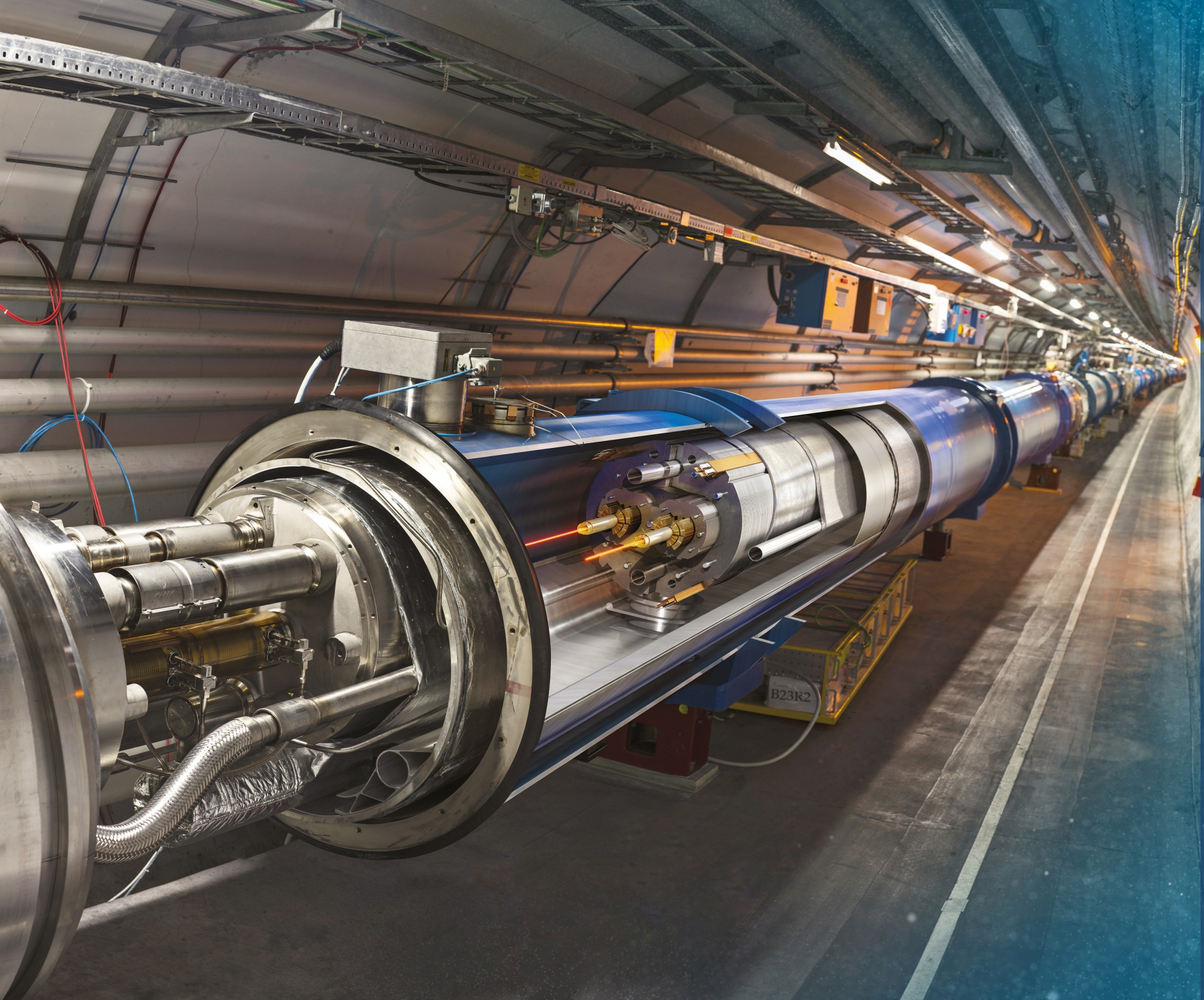


CERN is the world's biggest laboratory for particle physics.

Our goal is to understand the most fundamental particles and laws of the universe.

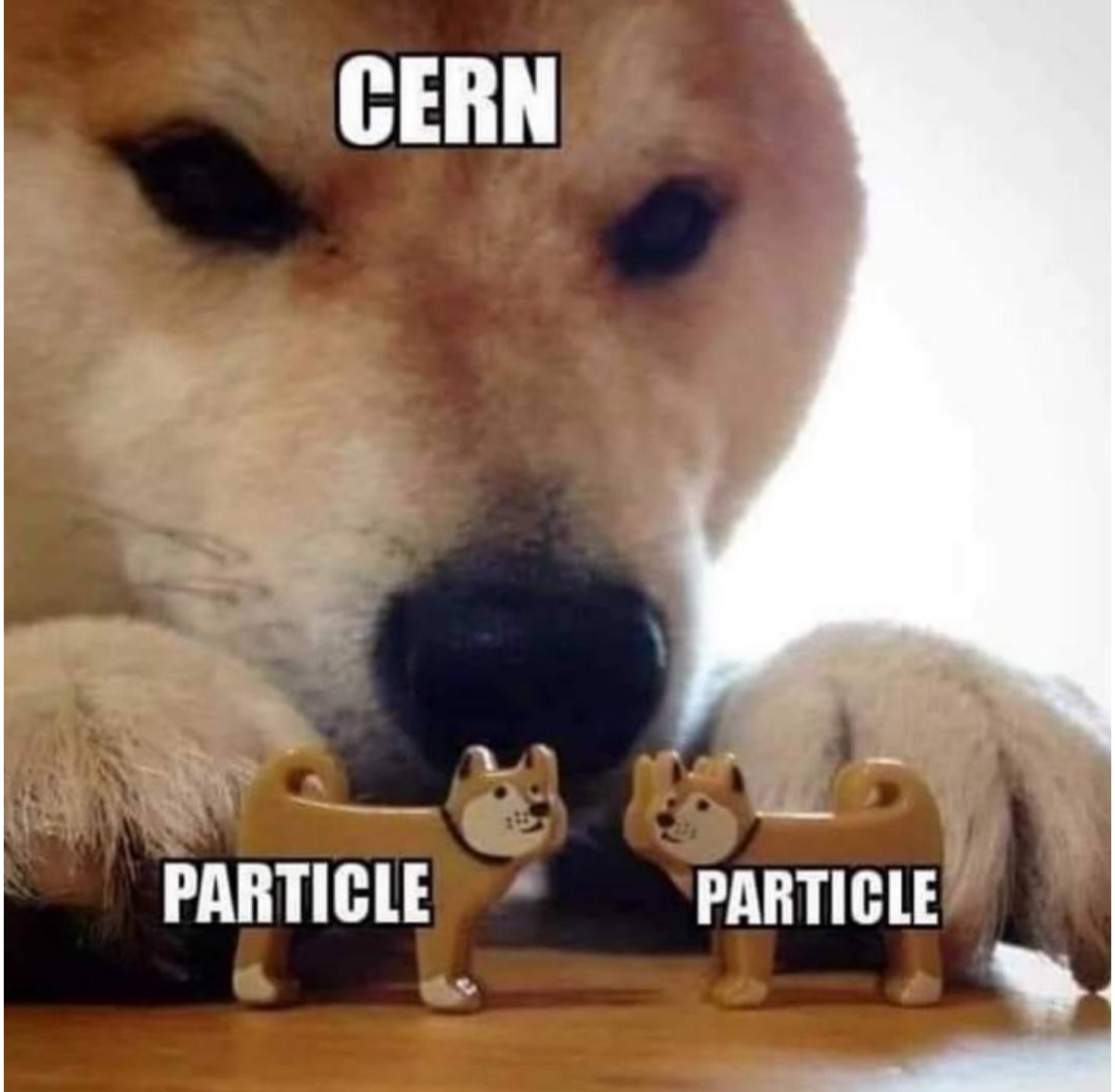


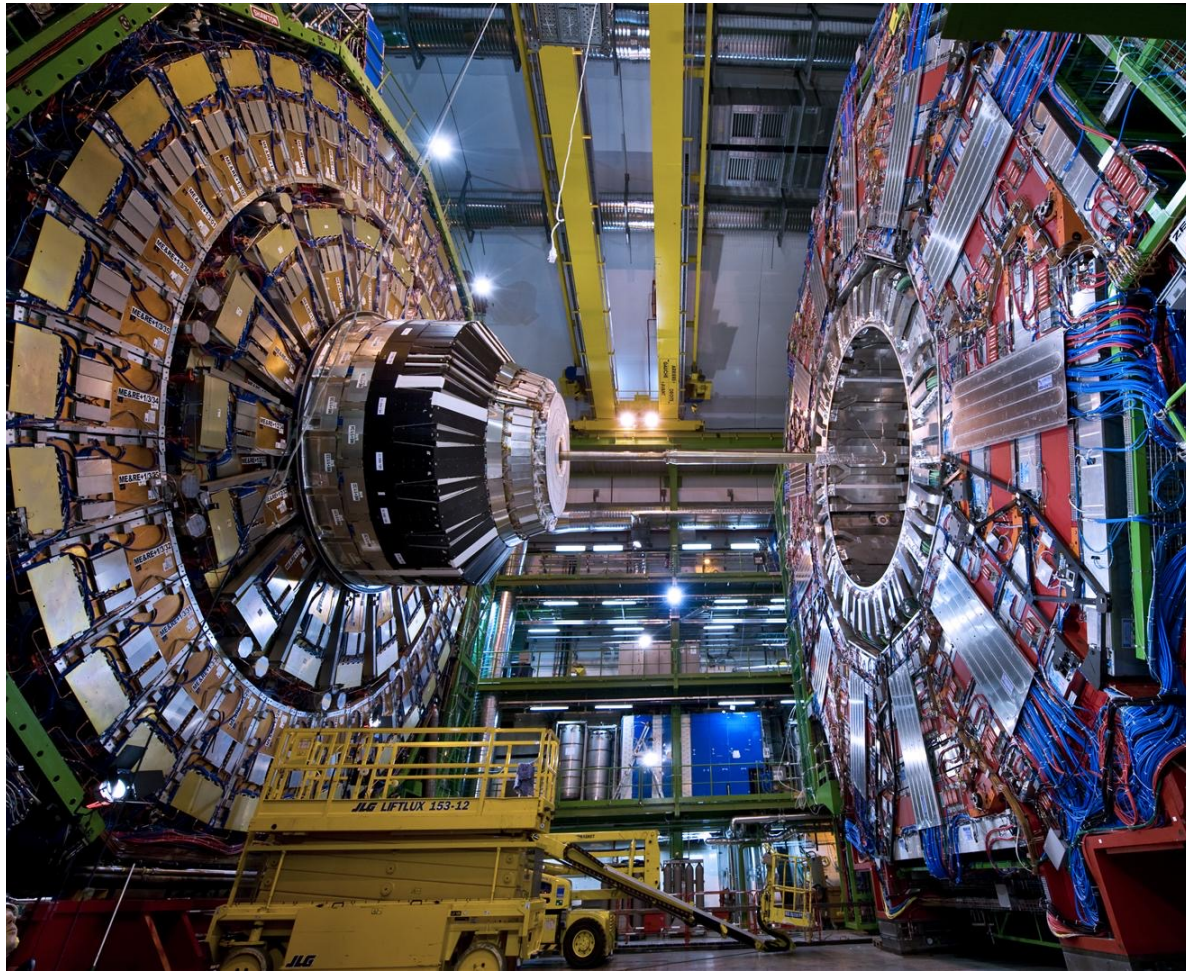
Large Hadron Collider (LHC)

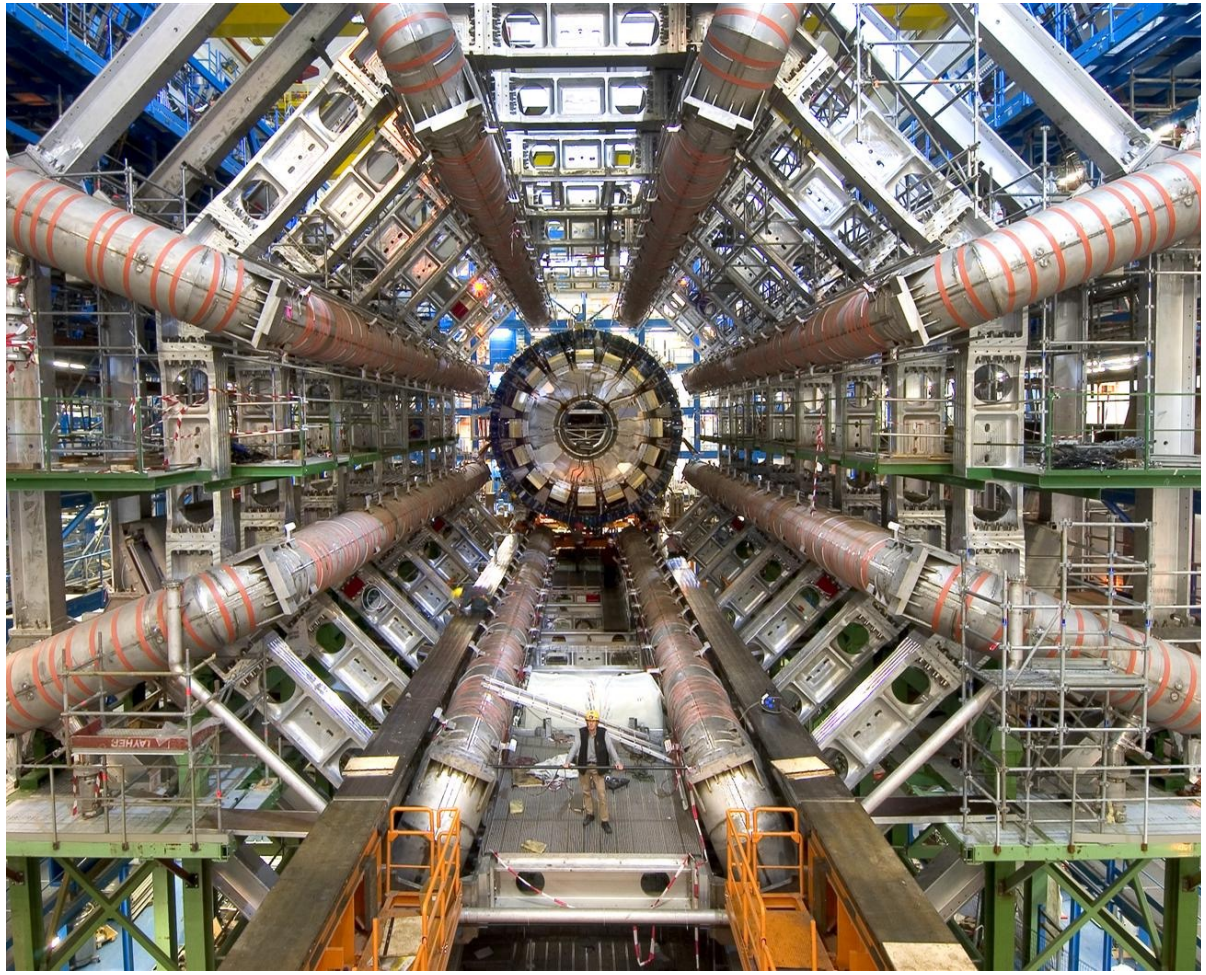


Large Hadron Collider (LHC)

- 27 km (17 mi) in circumference
- About 100 m (300 ft) underground
- Superconducting magnets steer the particles around the ring
- Particles are accelerated to close to the speed of light


CERN
PARTICLE
PARTICLE


JLG LIFTLUX 153-12
JLG



IT @ CERN

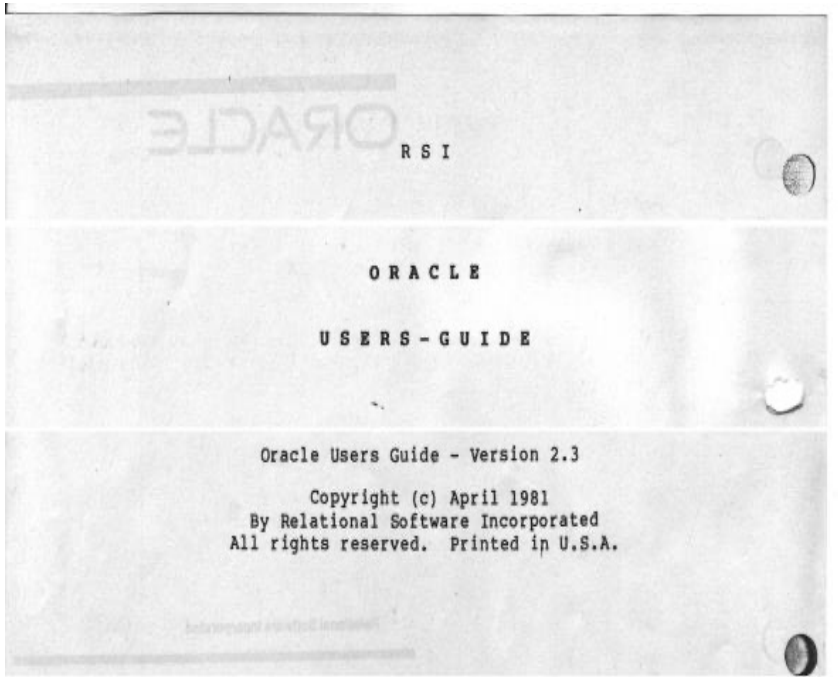
Databases at CERN

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)



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

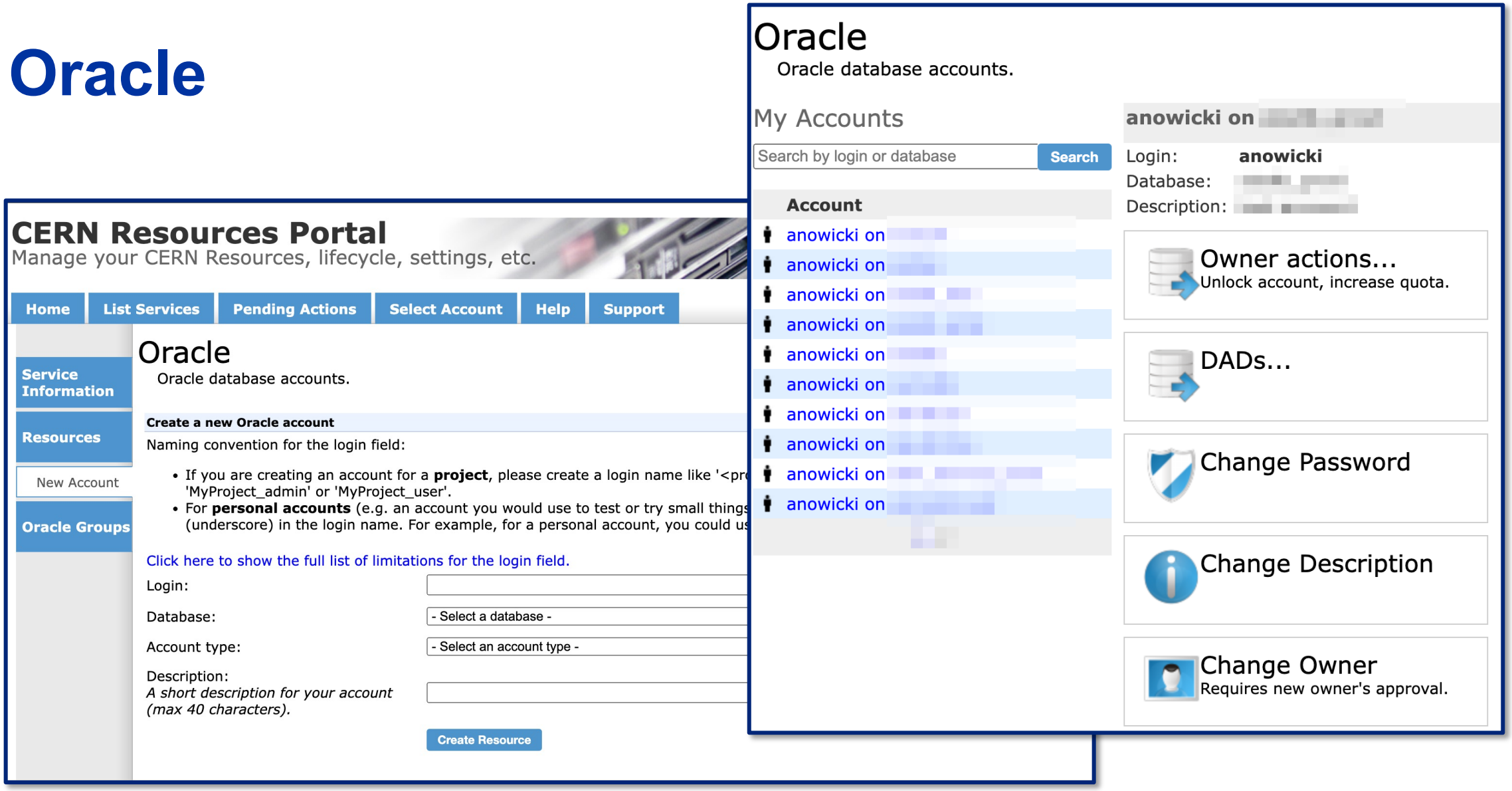
Each resource should have an owner, whom you can contact in case of problems.

We give two options:

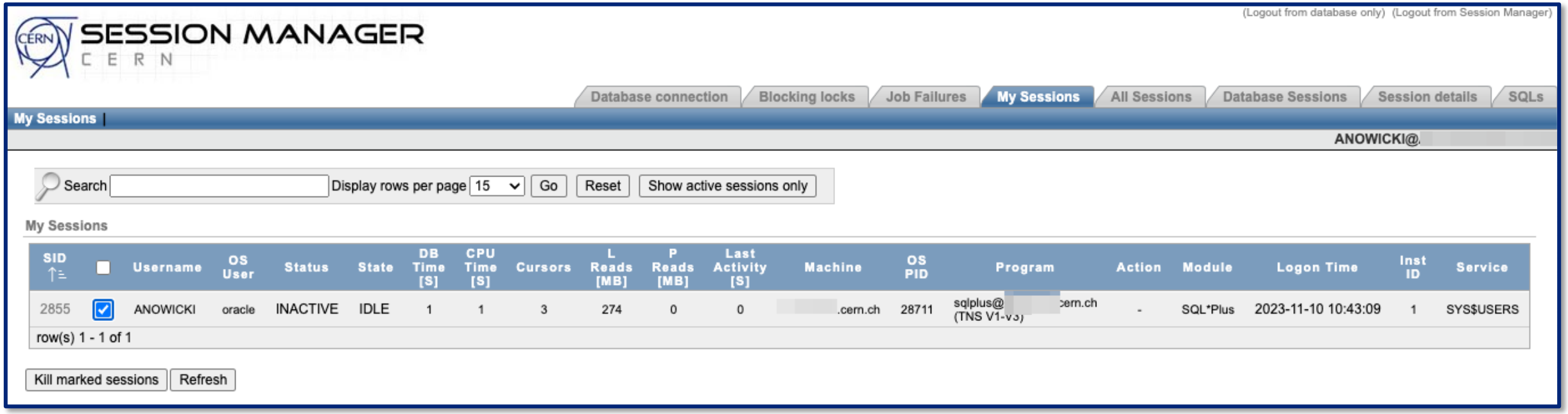
- **individual owner**
- **service owner (which might belong to a team, group, etc.)**

What do you do when people leave the organisation?

Oracle


CERN Resources Portal
Manage your CERN Resources, lifecycle, settings, etc.
Home
List Services
Pending Actions
Select Account
Help
Support
Service Information
Resources
New Account
Oracle Groups
Oracle
Oracle database accounts.
Create a new Oracle account
Naming convention for the login field:
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
Search by login or database
Search
Account
anowicki on
anowicki on
anowicki on
anowicki on
anowicki on
anowicki on
anowicki on
anowicki on
anowicki on
anowicki on
anowicki on
Login: anowicki
Database:
Description:
Owner actions...
Unlock account, increase quota.
DADs...
Change Password
Change Description
Change Owner
Requires new owner's approval.

Oracle


SESSION MANAGER
CERN
(Logout from database only) (Logout from Session Manager)
Database connection
Blocking locks
Job Failures
My Sessions
All Sessions
Database Sessions
Session details
SQLs
My Sessions |
ANOWICKI@
Search
Display rows per page 15
Go
Reset
Show active sessions only
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
.cern.ch
28711
sqlplus@
cern.ch
(TNS V1-V3)
-
SQL*Plus
2023-11-10 10:43:09
1
SYS$USERS
row(s) 1 - 1 of 1
Kill marked sessions
Refresh

Automate as much as possible

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

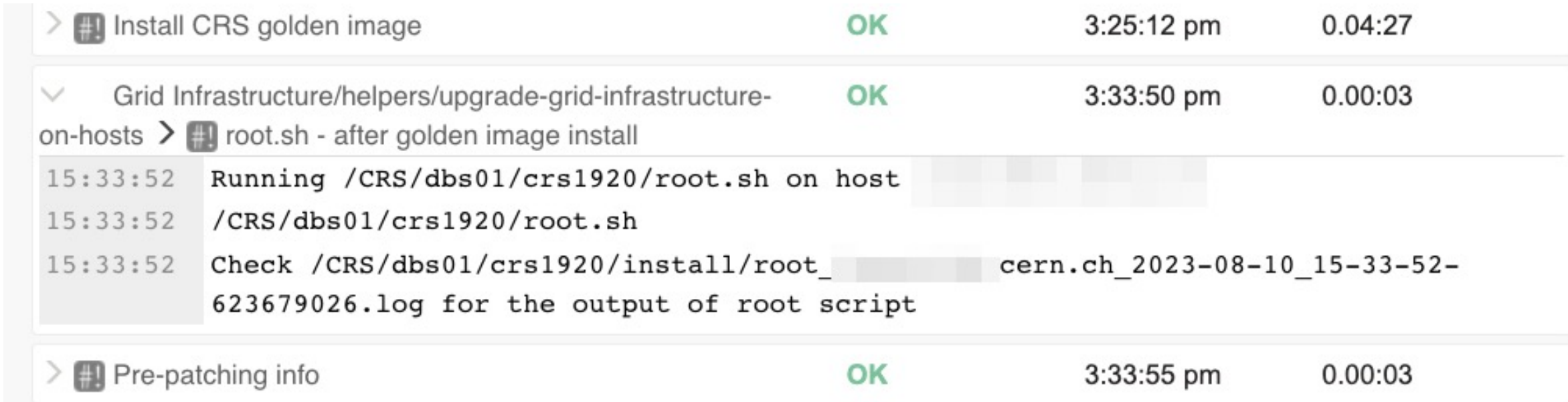
username

794f0e81

Run Again

Log Output »

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



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

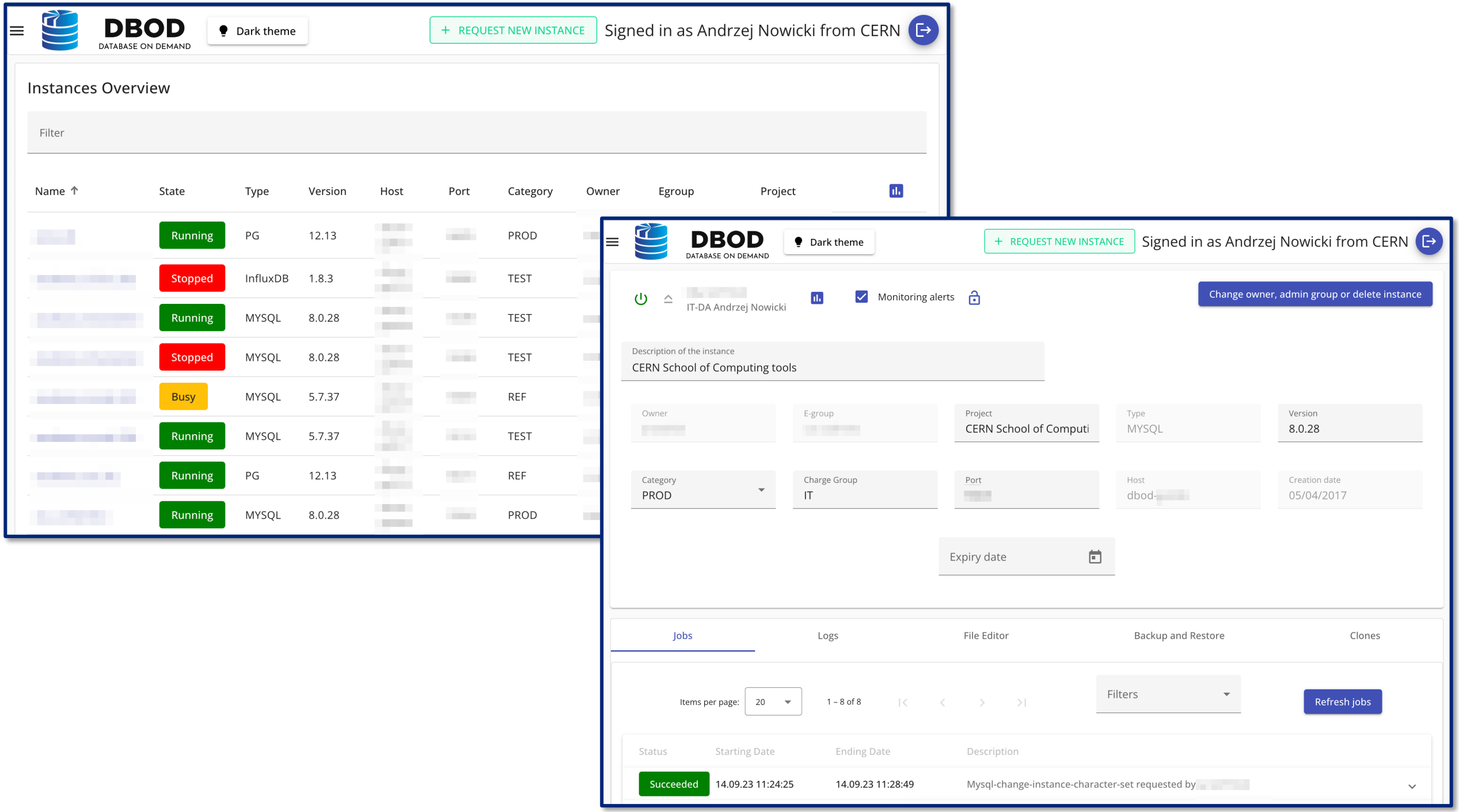
* Database type: Please select

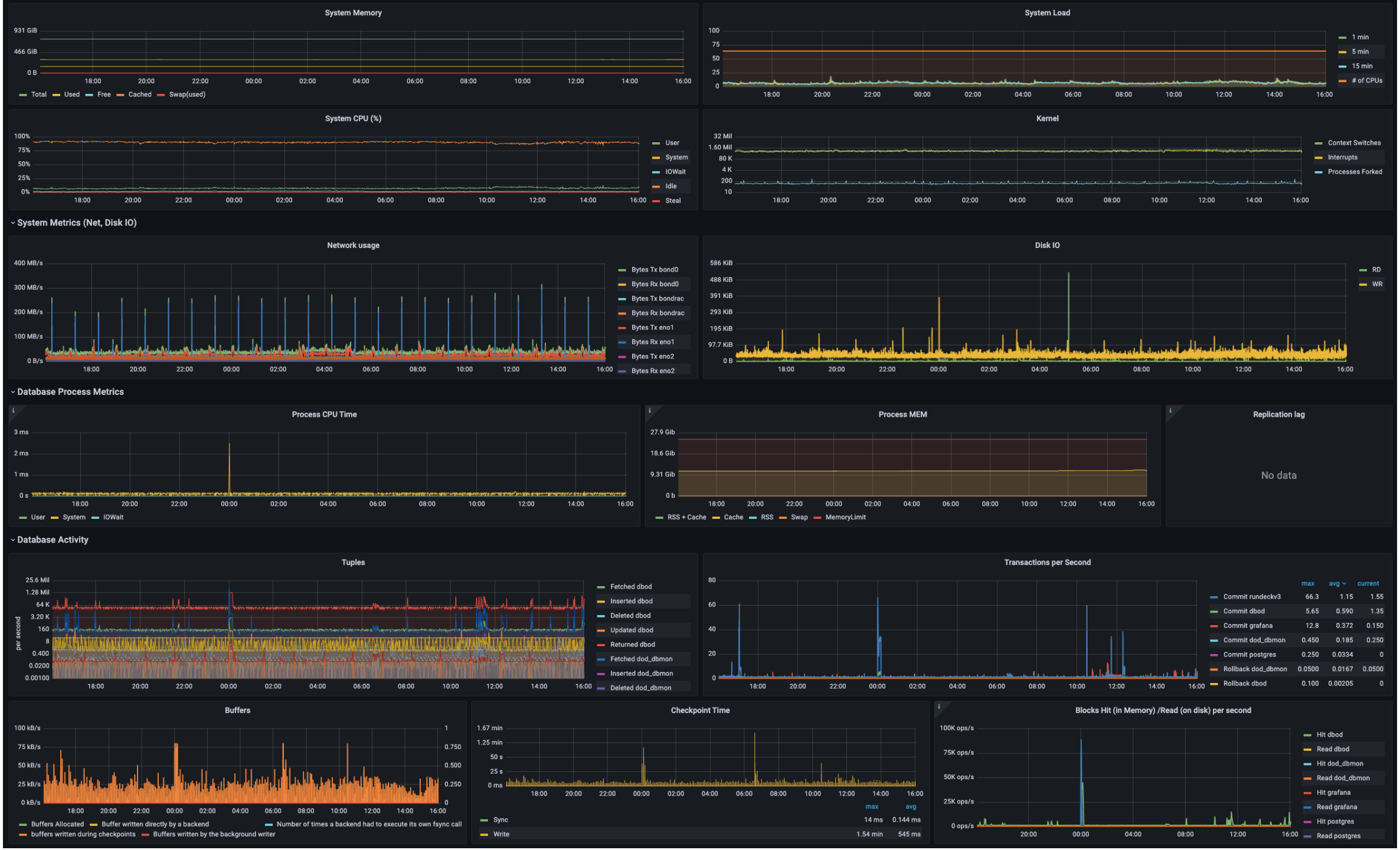
* Description: Please provide a description for your account. DB On Demand administrators will use this information to evaluate your request (max 200 characters).

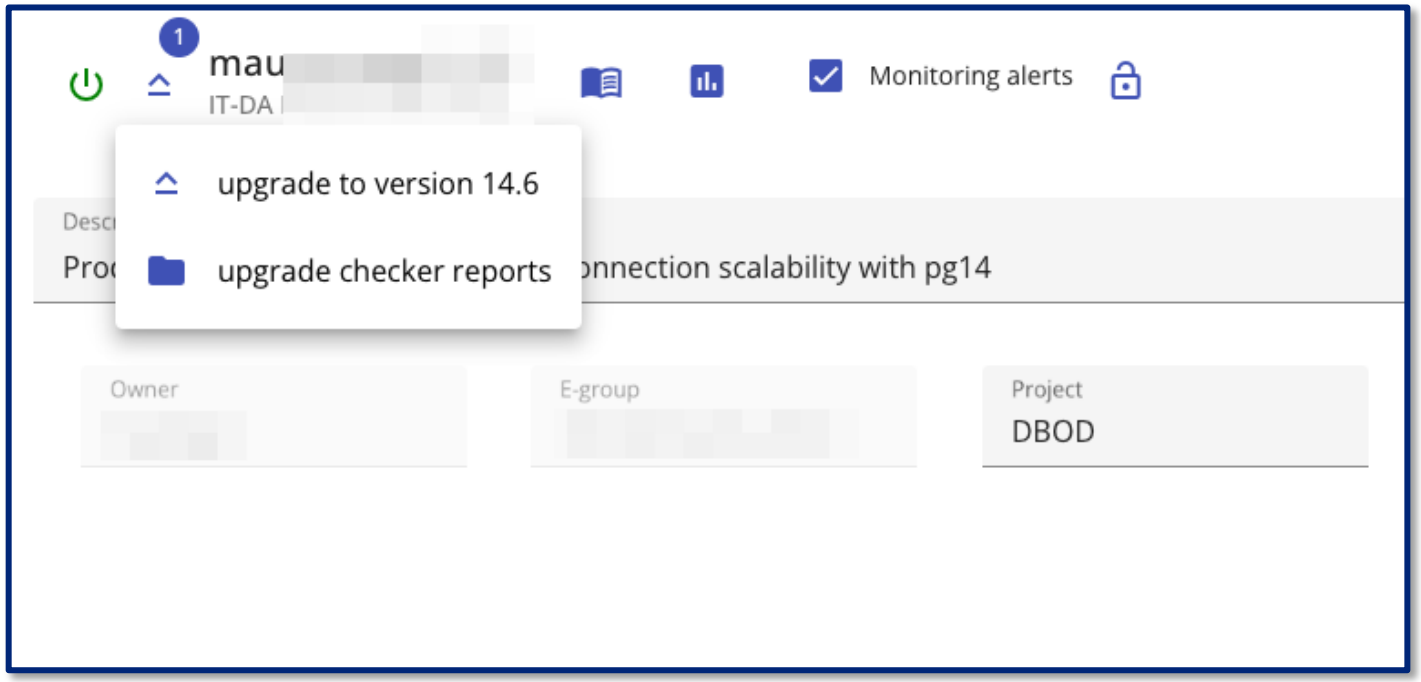
* Manifesto: By requesting a database, we consider you have fully read and accepted our manifesto.

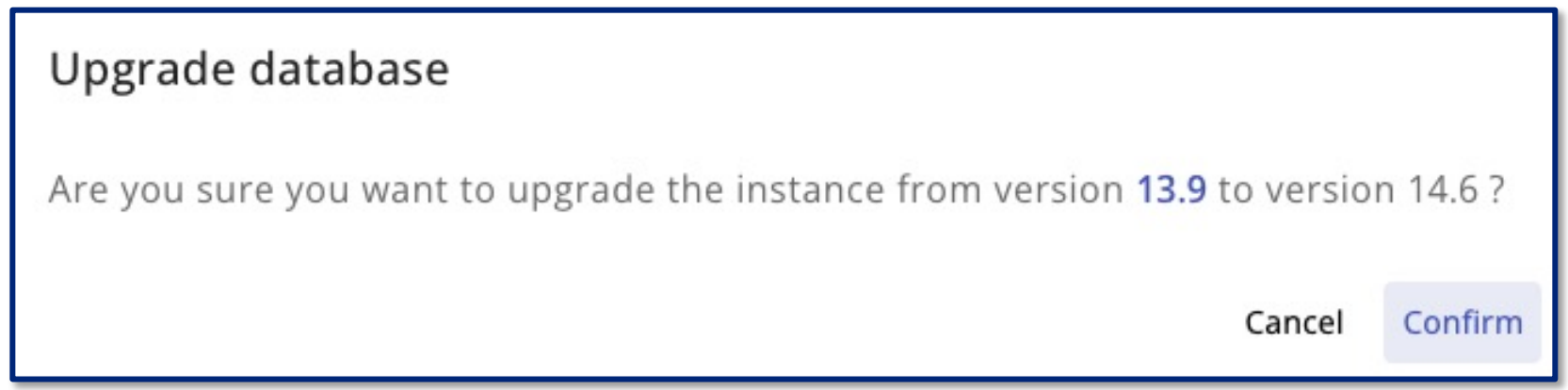
Note: The database will be available after verification and approval from the DBOD team.

Save | Cancel


DBOD
DATABASE ON DEMAND
Dark theme
+ REQUEST NEW INSTANCE
Signed in as Andrzej Nowicki from CERN
Instances Overview
Filter
Name
State
Type
Version
Host
Port
Category
Owner
Egroup
Project
Running PG 12.13 PROD
Stopped InfluxDB 1.8.3 TEST
Running MYSQL 8.0.28 TEST
Stopped MYSQL 8.0.28 TEST
Busy MYSQL 5.7.37 REF
Running MYSQL 5.7.37 TEST
Running PG 12.13 REF
Running MYSQL 8.0.28 PROD
DBOD
DATABASE ON DEMAND
Dark theme
+ REQUEST NEW INSTANCE
Signed in as Andrzej Nowicki from CERN
IT-DA Andrzej Nowicki
Monitoring alerts
Change owner, admin group or delete instance
Description of the instance
CERN School of Computing tools
Owner
E-group
Project
CERN School of Computi
Type
MYSQL
Version
8.0.28
Category
PROD
Charge Group
IT
Port
Host
dbod-
Creation date
05/04/2017
Expiry date
Jobs
Logs
File Editor
Backup and Restore
Clones
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


System Memory
931 GiB
466 GiB
0 B
Total Used Free Cached Swap(used)
System Load
1 min
5 min
15 min
# of CPUs
System CPU (%)
User
System
IOWait
Idle
Steal
Kernel
Context Switches
Interrupts
Processes Forked
System Metrics (Net, Disk IO)
Network usage
Bytes Tx bond0
Bytes Rx bond0
Bytes Tx bondrac
Bytes Rx bondrac
Bytes Tx eno1
Bytes Rx eno1
Bytes Tx eno2
Bytes Rx eno2
Disk IO
RD
WR
Database Process Metrics
Process CPU Time
User System IOWait
Process MEM
RSS + Cache Cache RSS Swap MemoryLimit
Replication lag
No data
Database Activity
Tuples
per second
Fetched dbod
Inserted dbod
Deleted dbod
Updated dbod
Returned dbod
Fetched dod_dbmon
Inserted dod_dbmon
Deleted dod_dbmon
Transactions per Second
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
Buffers
Buffers Allocated
Buffer written directly by a backend
Number of times a backend had to execute its own fsync call
buffers written during checkpoints
Buffers written by the background writer
Checkpoint Time
max avg
Sync 14 ms 0.144 ms
Write 1.54 min 545 ms
Blocks Hit (in Memory) /Read (on disk) per second
Hit dbod
Read dbod
Hit dod_dbmon
Read dod_dbmon
Hit grafana
Read grafana
Hit postgres
Read postgres


1
mau
IT-DA
Monitoring alerts
upgrade to version 14.6
upgrade checker reports
Desc
Proc
onnection scalability with pg14
Owner
E-group
Project
DBOD


Upgrade database
Are you sure you want to upgrade the instance from version 13.9 to version 14.6 ?
Cancel
Confirm

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)

Upgrade warning!
PostgreSQL is sensitive to glibc locale changes

RHEL/CentOS 9 locale changes sorting order (which is used to create indexes)

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

home.cern
CERN Accelerating science
Sign in
Directory
Drush Site-Install
HOME
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)
Copyright © 2022 CERN
CLOSE
DEVELOPMENT website

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**

Automate as much as possible – caveats

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?

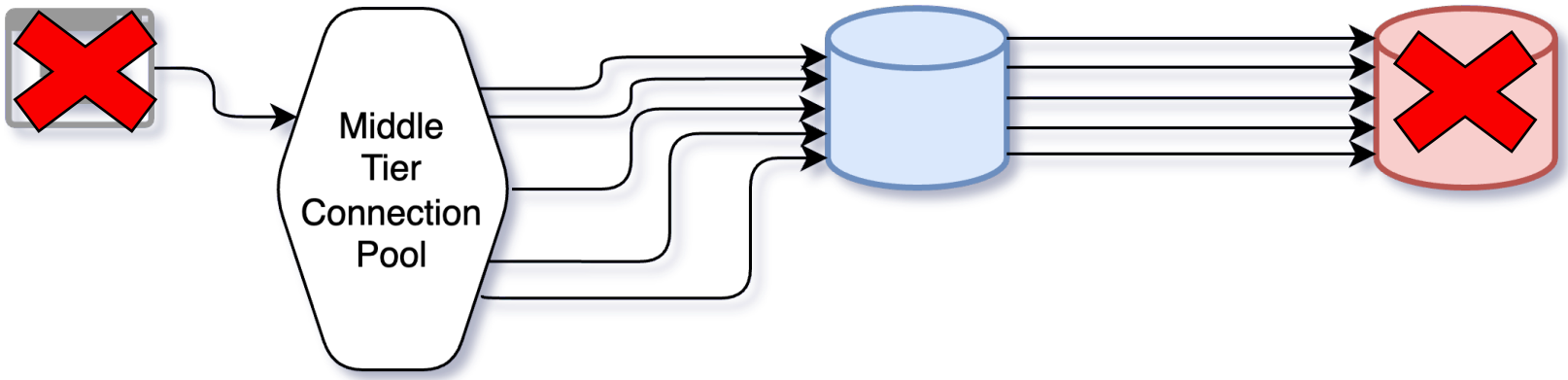
Will your automation still work?

Do you use DNS names to configure database replication?

- What if your DNS service is relying on your databases?
  Do you have `/etc/hosts` entries?

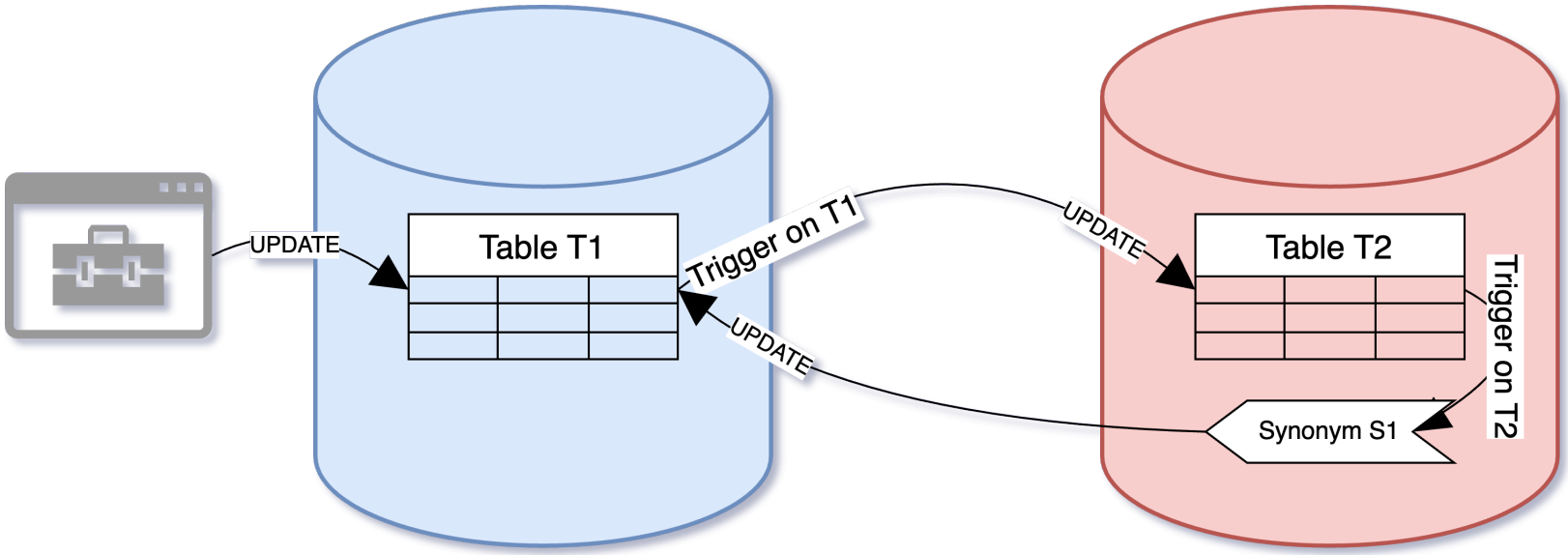
Database Links

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



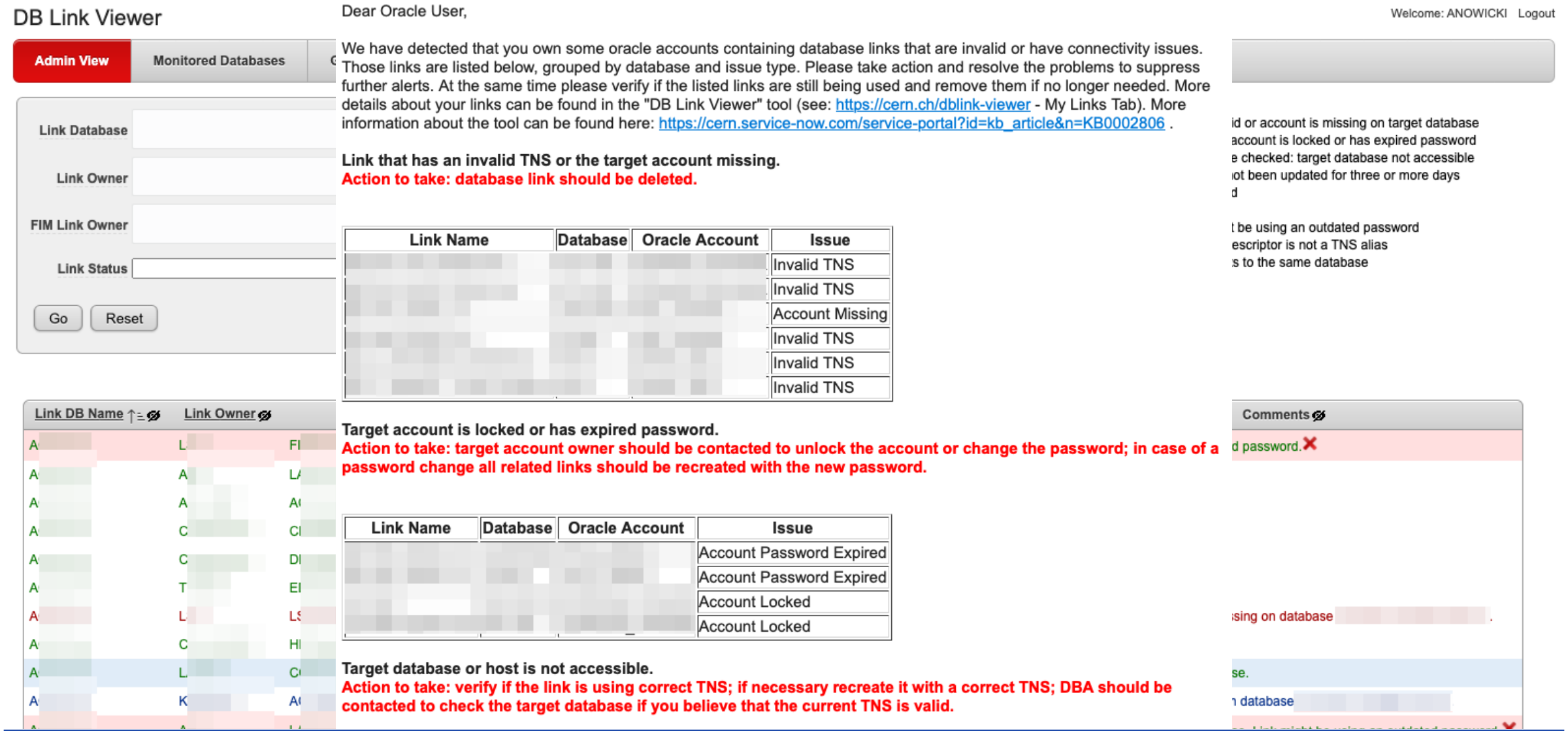
Database Links

Triggers on remote tables having triggers on remote tables?



Database Links

Database Links



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));

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
```

This invalidates the view

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:

- **Use Golden Images**
- **Recompile invalid objects after datapatch `(utlrp.sql)`**
  - It's done automatically only if you have <300 invalid objects
- **Look for documents of recommended patches:**
  - 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)
- **Follow https://mikedietrichde.com/**

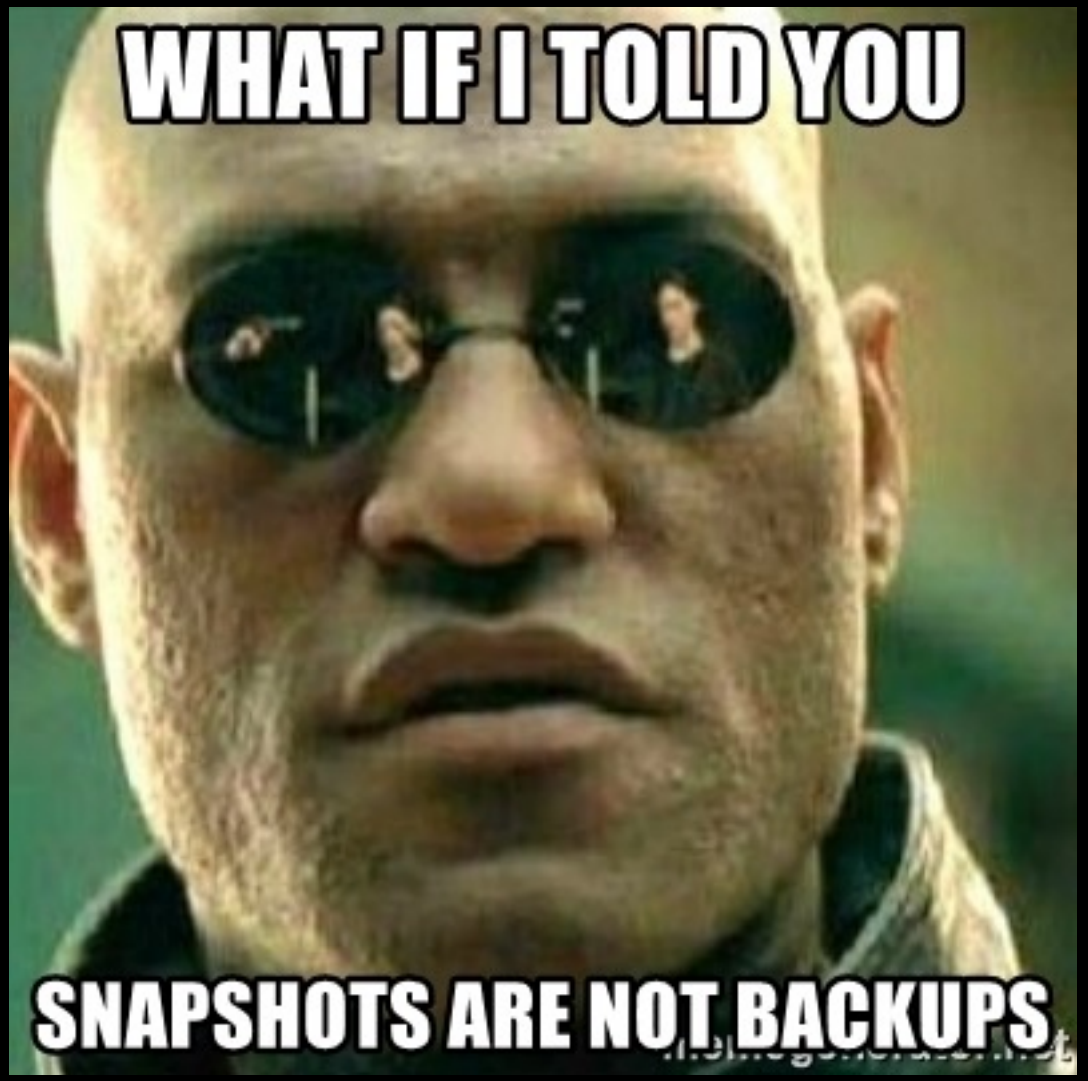
Backup

DOs:

- **Use RMAN catalog**
- **Snapshot & backup combination**
- **Automatic recoveries or recovery environment**
- **Consider immutable backups**

DON'Ts:

- **Snapshots ARE NOT backup**


WHAT IF I TOLD YOU
SNAPSHOTS ARE NOT BACKUPS

VLDBs

DOs:

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

VLDB and Partitioning Guide

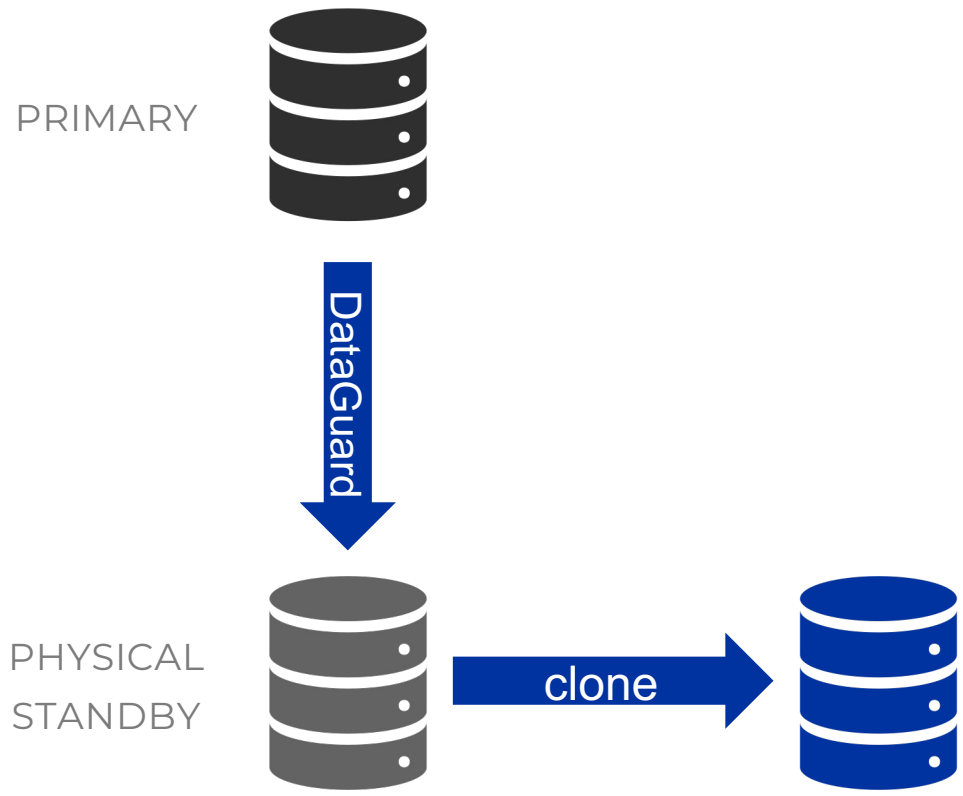
Data Pump automation

```
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');
```

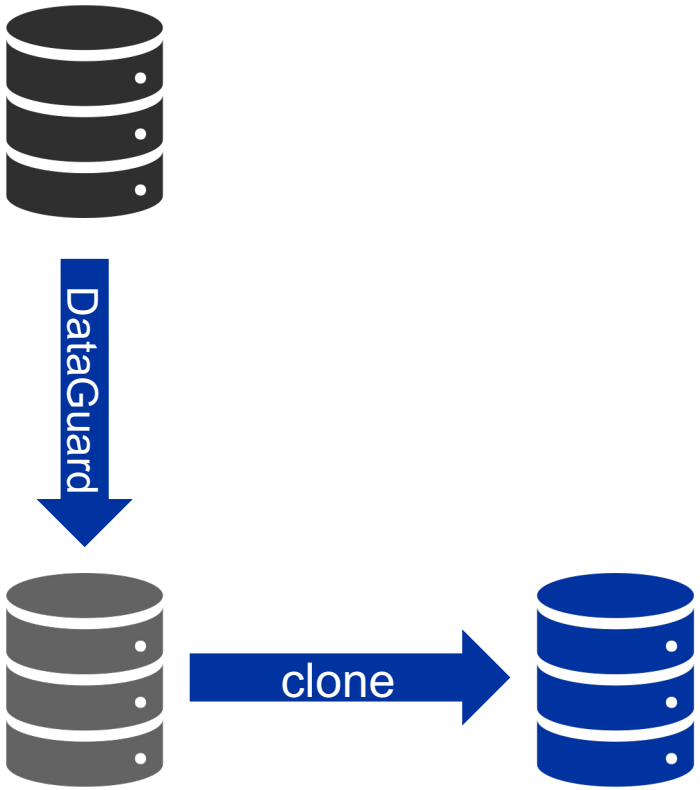
Database Clones

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

1. *When you stop apply, you get consistent datafiles*
2. *Create a filesystem snapshot*
3. *Use `clonedb.pl` & `dbms_dnfs.clonedb_renamefile`*

Database Clones

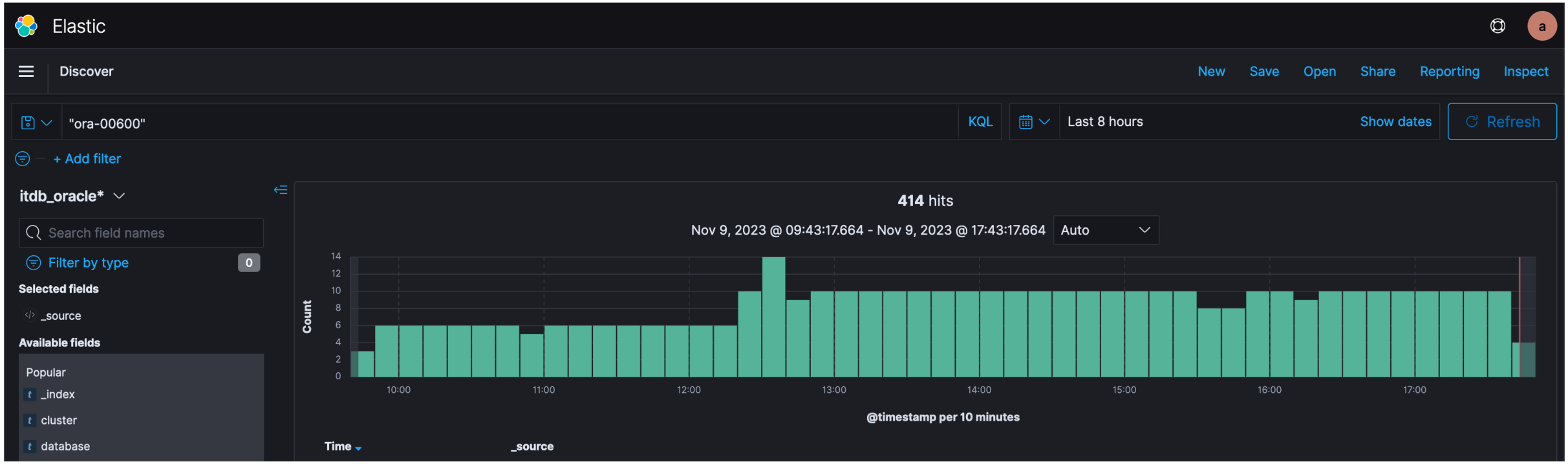
Oracle Multitenant Pluggable Database Snapshot Cloning: Use Cases and Supported Platforms (Doc ID 1597027.1)

There's also possibility of:

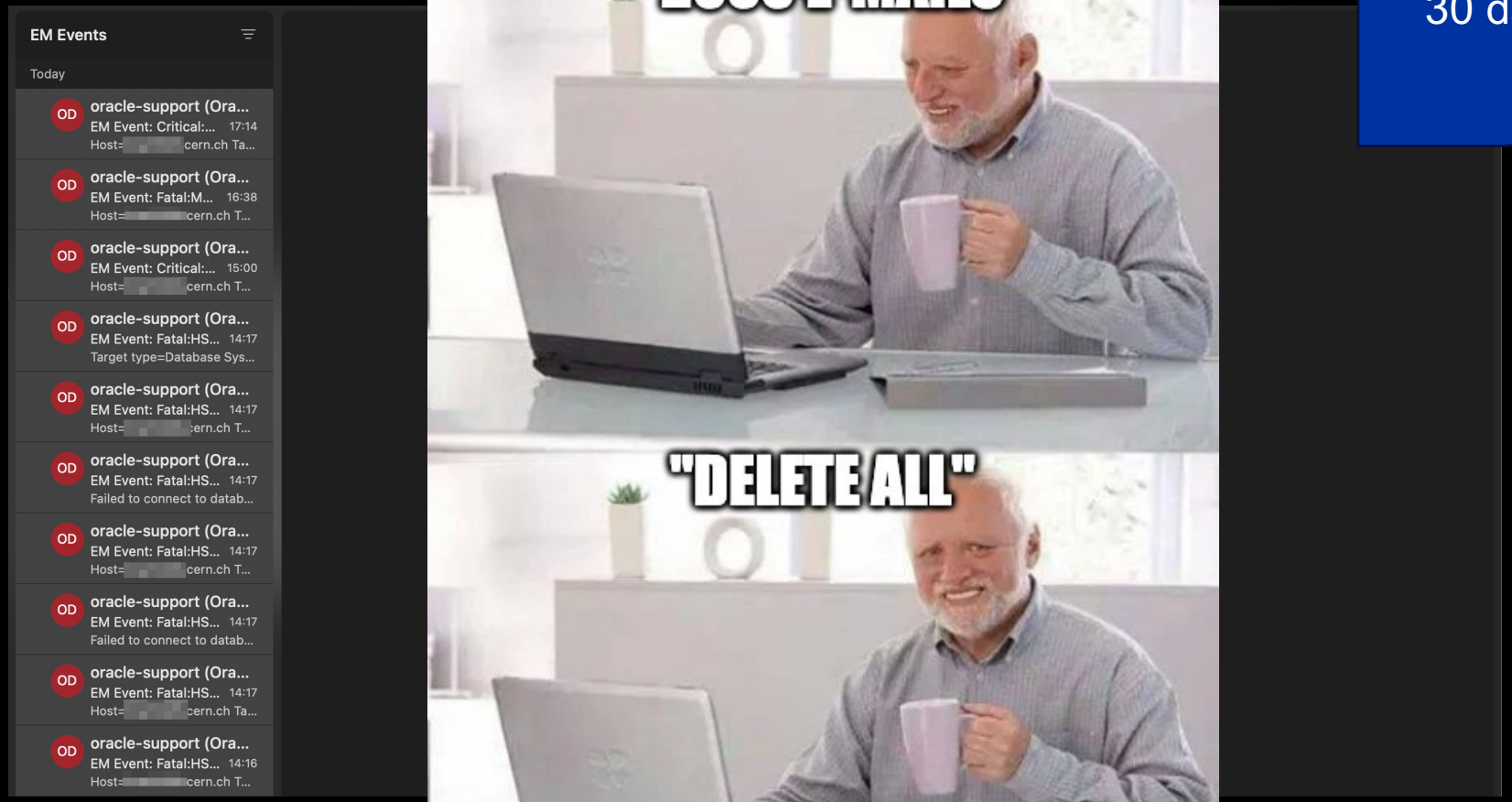
Example for Cloning PDB from NON-CDB via dblink (Doc ID 1928653.1)

Centralised logging

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



DON’T do alerting via e-mail



30 days

Things to consider in the future

AHF Insights – report on settings, configuration, best practices

See the Cloud World AHF presentation by Sandesh Rao

Database Resident Connection Pooling

Due to a raising problem with kubernetes deployments creating several connections pools for the same app

Come visit CERN!
https://visit.cern

Thank you !

andrzejnowicki

andrzej.nowicki@cern.ch

www.andrzejnowicki.pl

CONFERENCE SPONSORS