pg_upgrade

Daniel Gustafsson
Pivotal

dgustafsson@pivotal.io @d_gustafsson

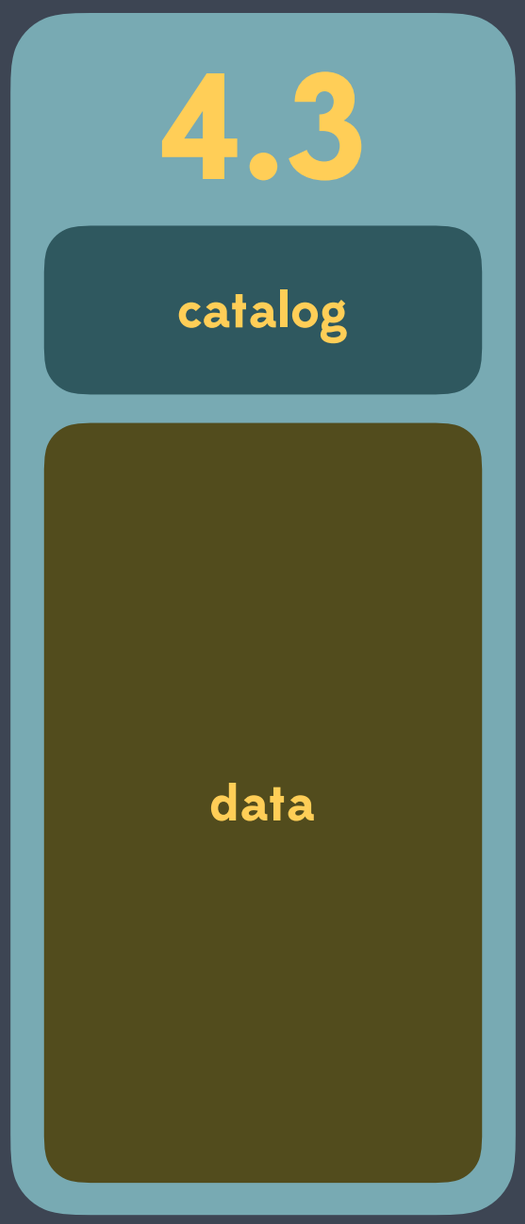
Previously only dump/restore for major version

On-disk format rarely (never) change; only catalog needs to be upgraded

Binary upgrade for major version upgrades

Optionally in-place for data

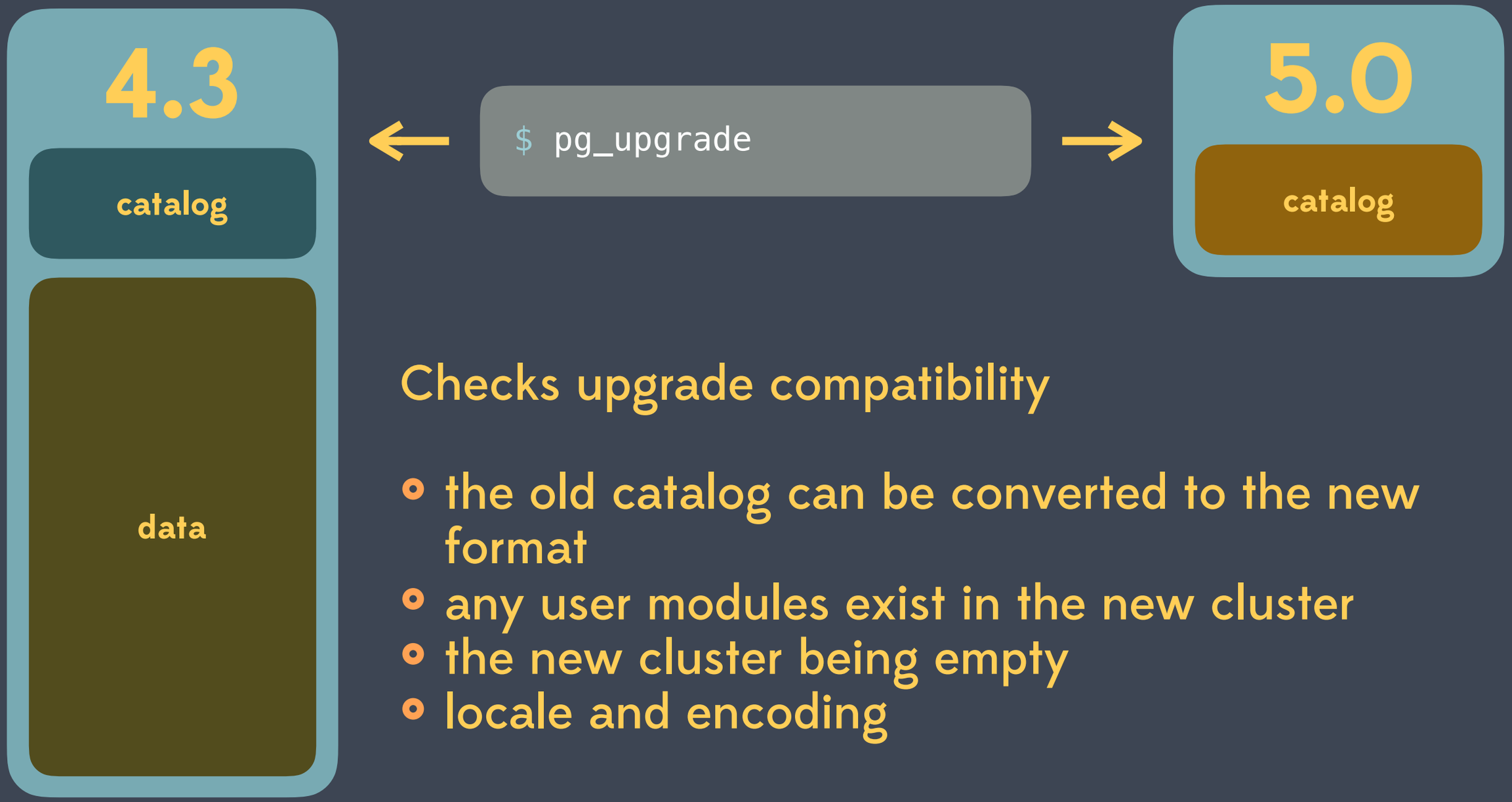
0. Pre-install steps



Before running pg_upgrade

- A fresh initdb (gpinitsystem) 5.0 cluster without user data/relations
- New and old cluster running with same encoding and locale
- Both clusters turned off, logins during upgrade prohibited

1. Check clusters



Checks upgrade compatibility

- the old catalog can be converted to the new format
- any user modules exist in the new cluster
- the new cluster being empty
- locale and encoding

2. Dump schema using binary upgrade


```
$ pg_upgrade
```

```
$ pg_dumpall --binary-upgrade --schema
```

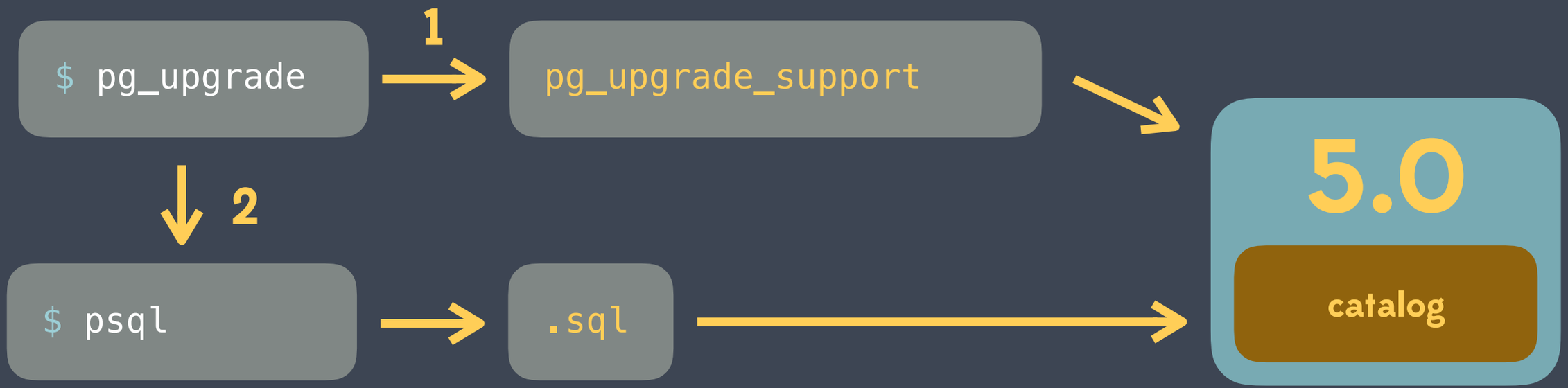
```
.sql
```

Injects OID assignment before DDL in dump file

```
-- For binary upgrade, must preserve pg_extprotocol oid
SELECT binary_upgrade.preassign_extprotocol_oid(
  ‘17157’::pg_extprotocol.oid,
  ‘demoprot’::text
);

CREATE TRUSTED PROTOCOL demoprot (
  readfunc = ‘read_from_file',
  writefunc = ‘write_to_file'
);
```

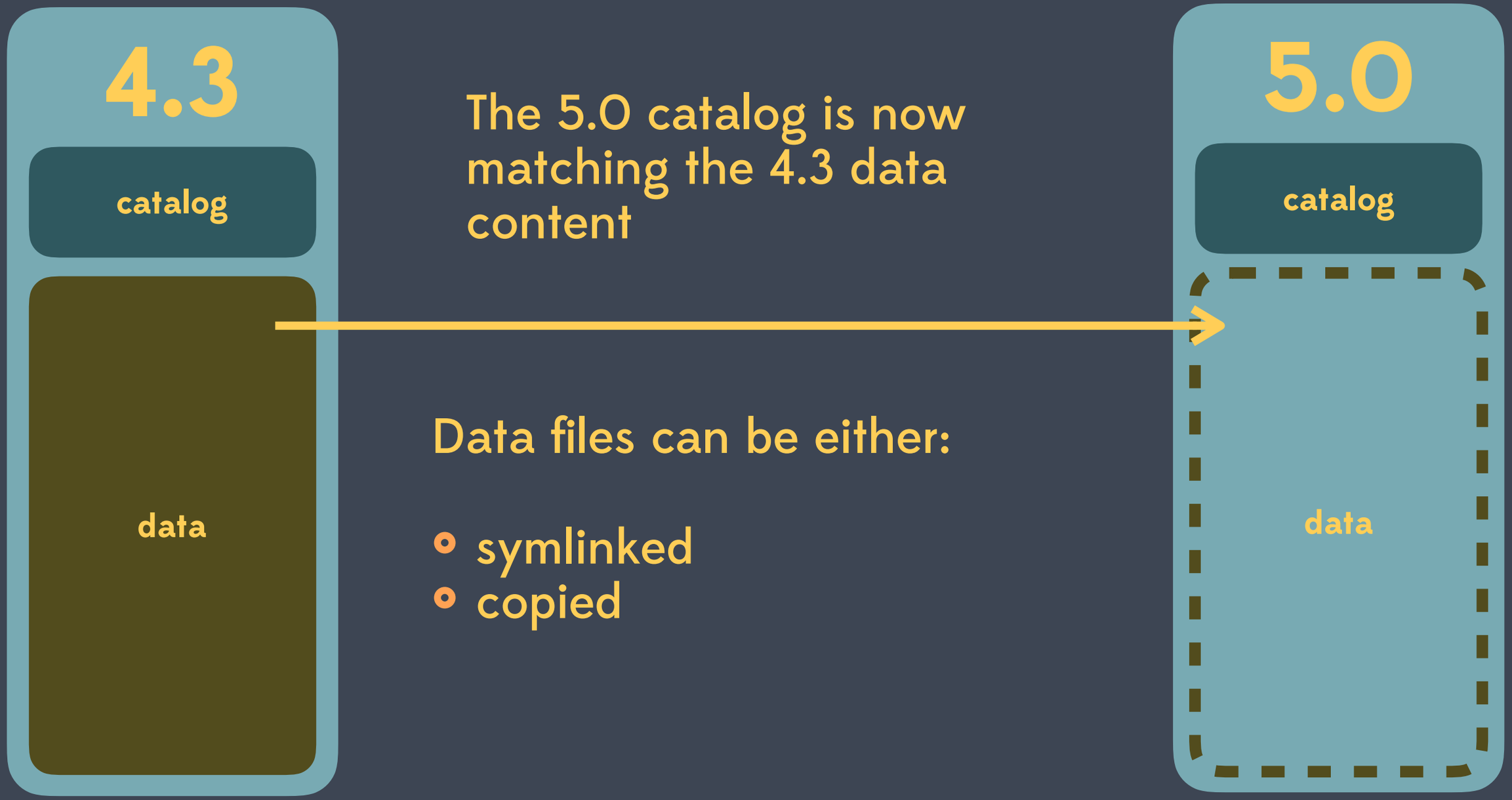
3. Restore schema



Schema restore via psql

- each segment is connected to in utility mode with binary-upgrade
- pg_upgrade_support functions are installed
- during the restoration of the dump, the captured OIDs are injected into the preassigned_oids list
- DDL CREATE will pick up the right OID

4. Transfer data



GPDB is not as simple as upstream since we're going from 8.2

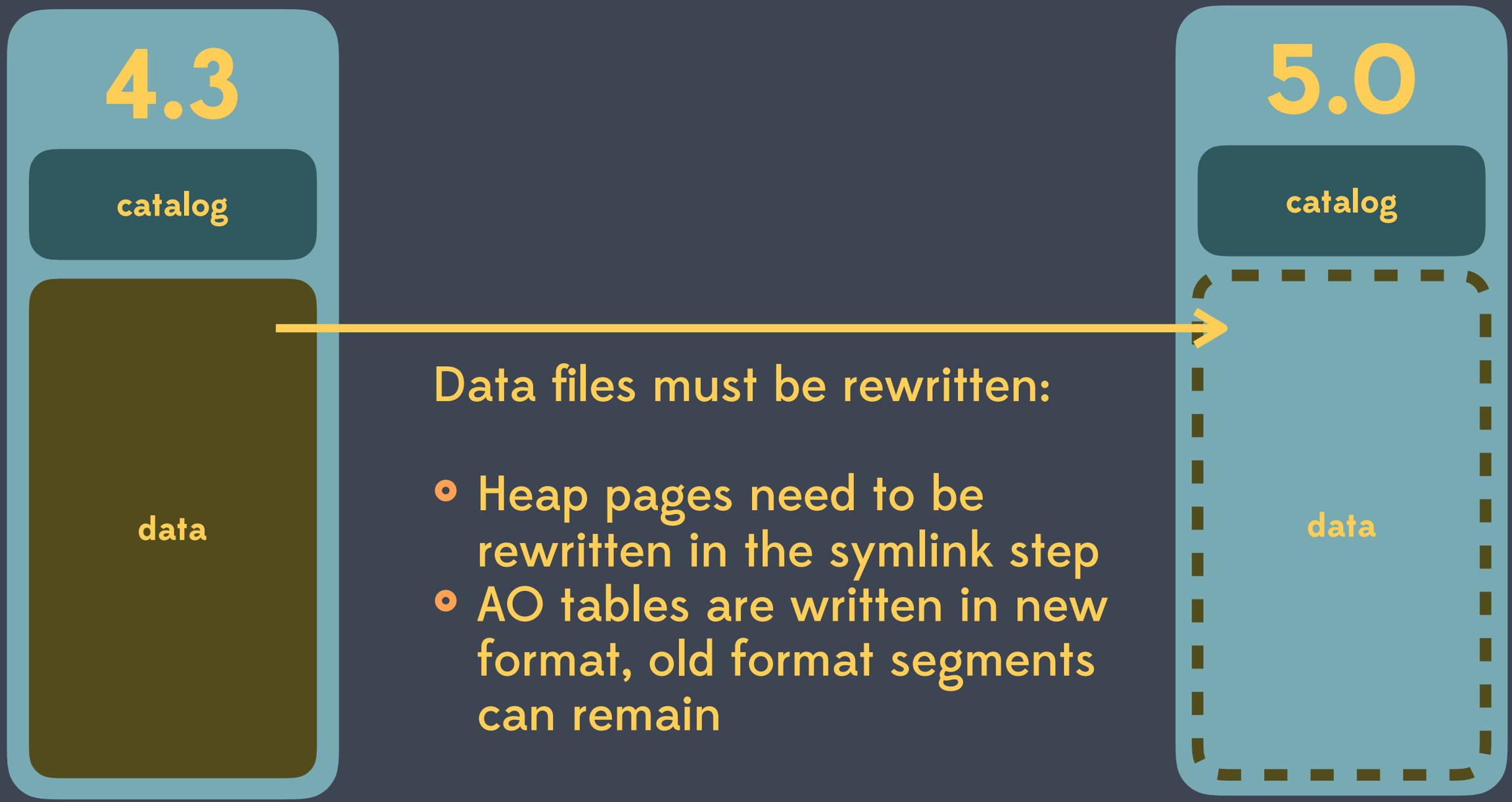
AO/AOCS tables

On-disk format change (last one..)

OID synchronisation

Money datatype 💀

5. GPDB on-disk format change



- Heap pages need to be rewritten in the symlink step
- AO tables are written in new format, old format segments can remain

GPDB had to be improved for pg_upgrade to work

OID synchronisation (#1263)

Per-segment AO version (#995)


SHOW ME
THE CODE

src/backend/catalog/**oid_dispatch.c**

New functionality for dispatching OID from QD to QE for normal DDL

QD sets the dispatch list which populates the preassigned list in QE's

In binary upgrade mode, the preassigned list can be populated directly without QD

`AddPreassignedOidFromBinaryUpgrade()`

contrib/pg_upgrade_support/**pg_upgrade_support.c**

Stored procedures for accessing the preassigned OID list

```
Datum
preassign_relation_oid(PG_FUNCTION_ARGS)
{
   Did     reloid = PG_GETARG_OID(0);
   char   *relname = GET_STR(PG_GETARG_TEXT_P(1));
   Did     relnamespace = PG_GETARG_OID(2);

   if (Gp_role == GP_ROLE_UTILITY)
   {
      AddPreassignedOidFromBinaryUpgrade(reloid, RelationRelationId,
                                         relname, relnamespace,
                                         InvalidOid, InvalidOid);
   }

   PG_RETURN_VOID();
}
```

src/bin/pg_dump/**pg_dump.c**

Dumping objects with OID injection

```
if (binary_upgrade)
    binary_upgrade_preassign_language_oid(defqry, plang->dobj.catId.oid,
                                          plang->dobj.name);

appendPQExpBuffer(defqry, "CREATE %sPROCEDURAL LANGUAGE %s",
                  (useParams && plang->lanpltrusted) ? "TRUSTED " : "",
                  qlanname);
```

```
static void
binary_upgrade_preassign_language_oid(PQExpBuffer upgrade_buffer,
                                      Oid langoid, char *lanname)
{
    appendPQExpBuffer(upgrade_buffer,
        "\n-- For binary upgrade, preserve pg_language oid\n");
    appendPQExpBuffer(upgrade_buffer,
        "SELECT binary_upgrade.preassign_language_oid("
        "    '%u'::pg_catalog.oid, "
        "    '%s'::text);\n",
        langoid, lanname);
}
```

contrib/pg_upgrade/*.c

pg_upgrade back ported from 9.0.23

- check.c checks for cluster compatibility
- pg_upgrade.c main logic code path
- version_old_gpdb4.c check GPDB4 specifics in the old cluster

contrib/pg_upgrade/**gpdb4_heap_convert.c**

Functionality for converting heap tables from 4.3 to 5.0 format

- **pd_prune_xid** added to page header
- Line pointer flags changed
- **HEAP_COMPRESSED** flag removed
- On-disk format for **NUMERIC** changed

.. and then some more

```
62 files changed, 9692 insertions(+), 134 deletions(-)
```

Mostly done, some nits remain..

Toasted numerics needs to untested during conversion

AO/AOCS tables where all segments are used up (pg_upgrade will warn)

Array types of base relations

Tests and pipeline

Lots of hacking by Heikki (mainly), Dave and me. Patch in PR #1340

Easy hacking/testing:

1. clone and build 4.3
2. make cluster
3. create some tables
4. make cluster in a 5.0 tree
5. tar up 5.0 tree as backup
6. pg_upgrade