PGConf.Russia 2018

Российская конференция по PostgreSQL

Подключаемое хранилище для больших объектов в PostgreSQL

Валерий Косарев
RedSys

PostgreSQL

Предпосылки

PostgreSQL

ceph

REDSYS

Предпосылки

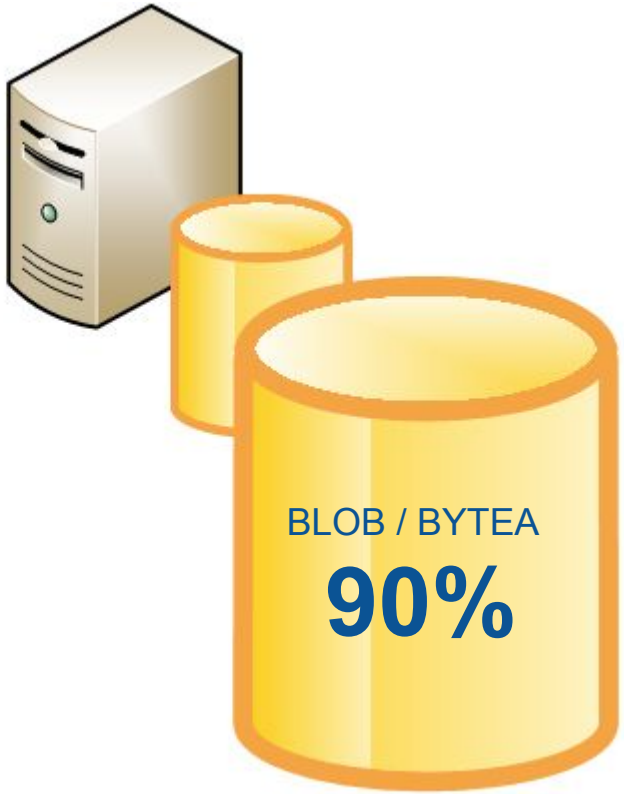
BLOB / BYTEA

PostgreSQL

ceph

REDSYS

Предпосылки

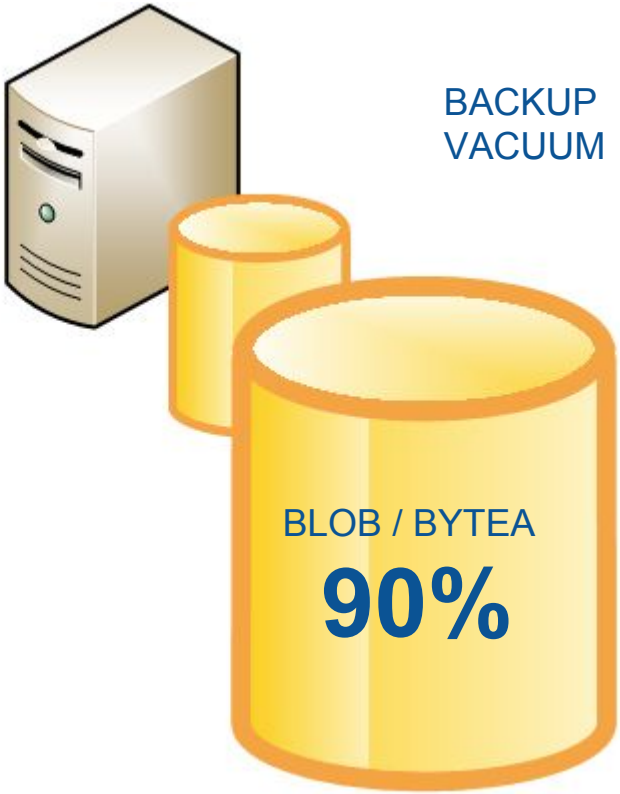


PostgreSQL

ceph

REDSYS

Предпосылки

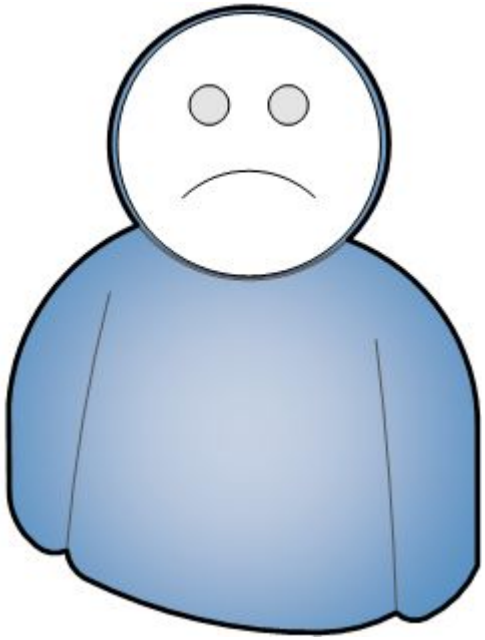


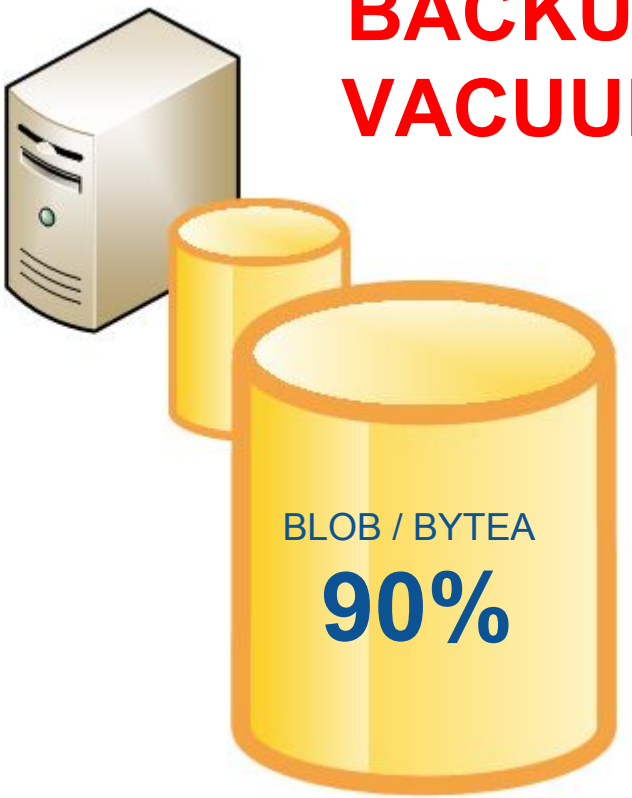
PostgreSQL

ceph

REDSYS

Предпосылки



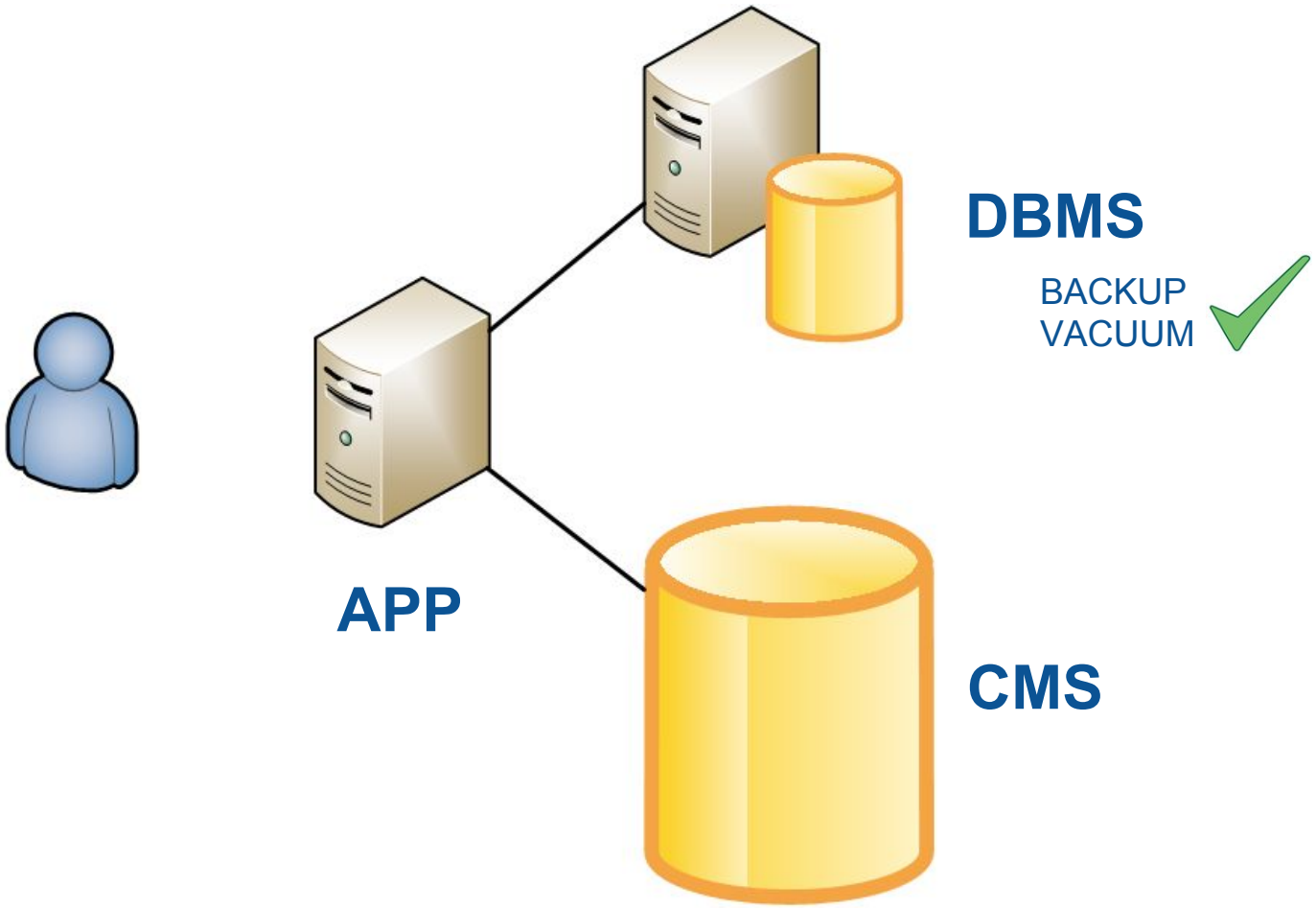


PostgreSQL

ceph

REDSYS

Предпосылки

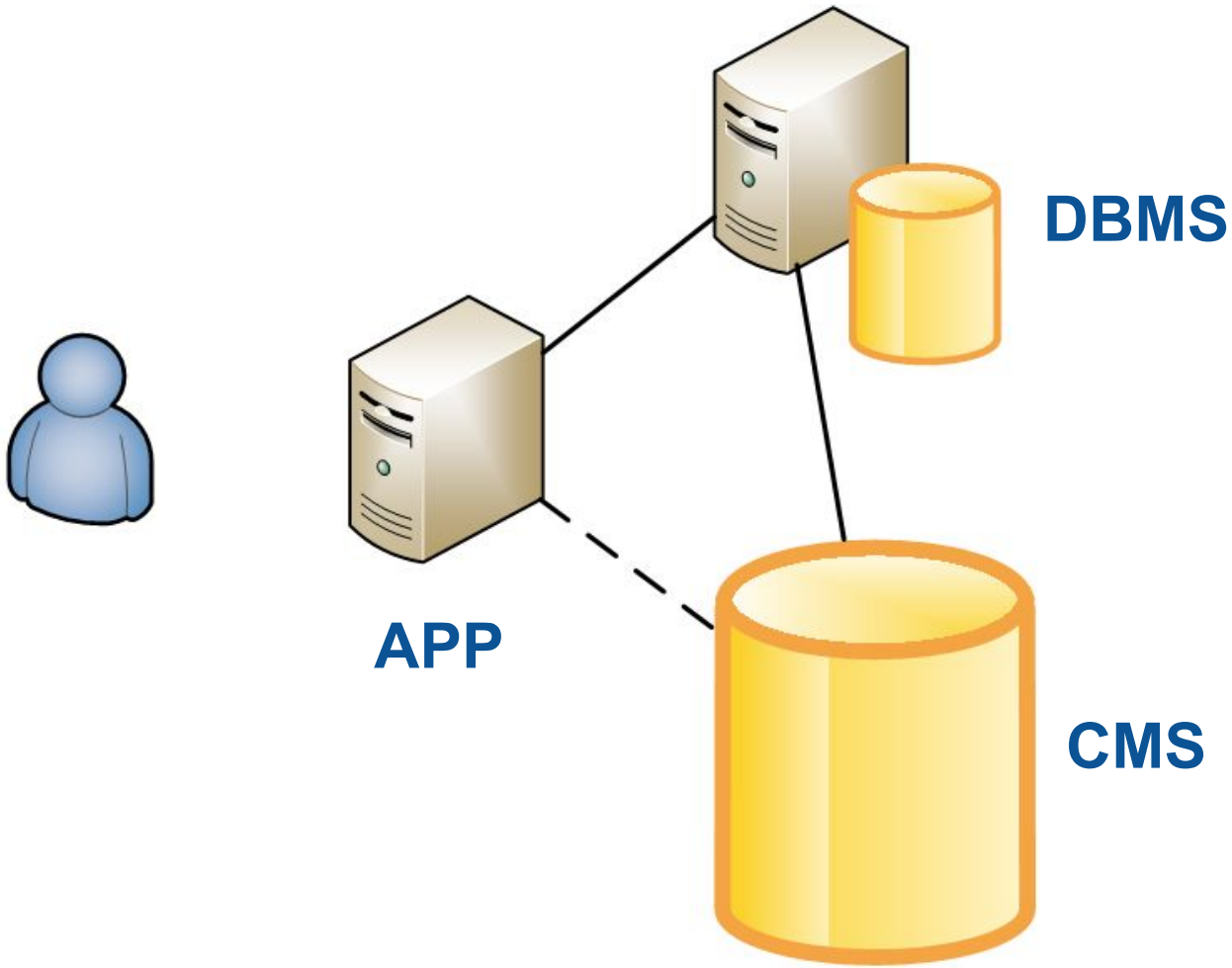


PostgreSQL

ceph

REDSYS

Предпосылки

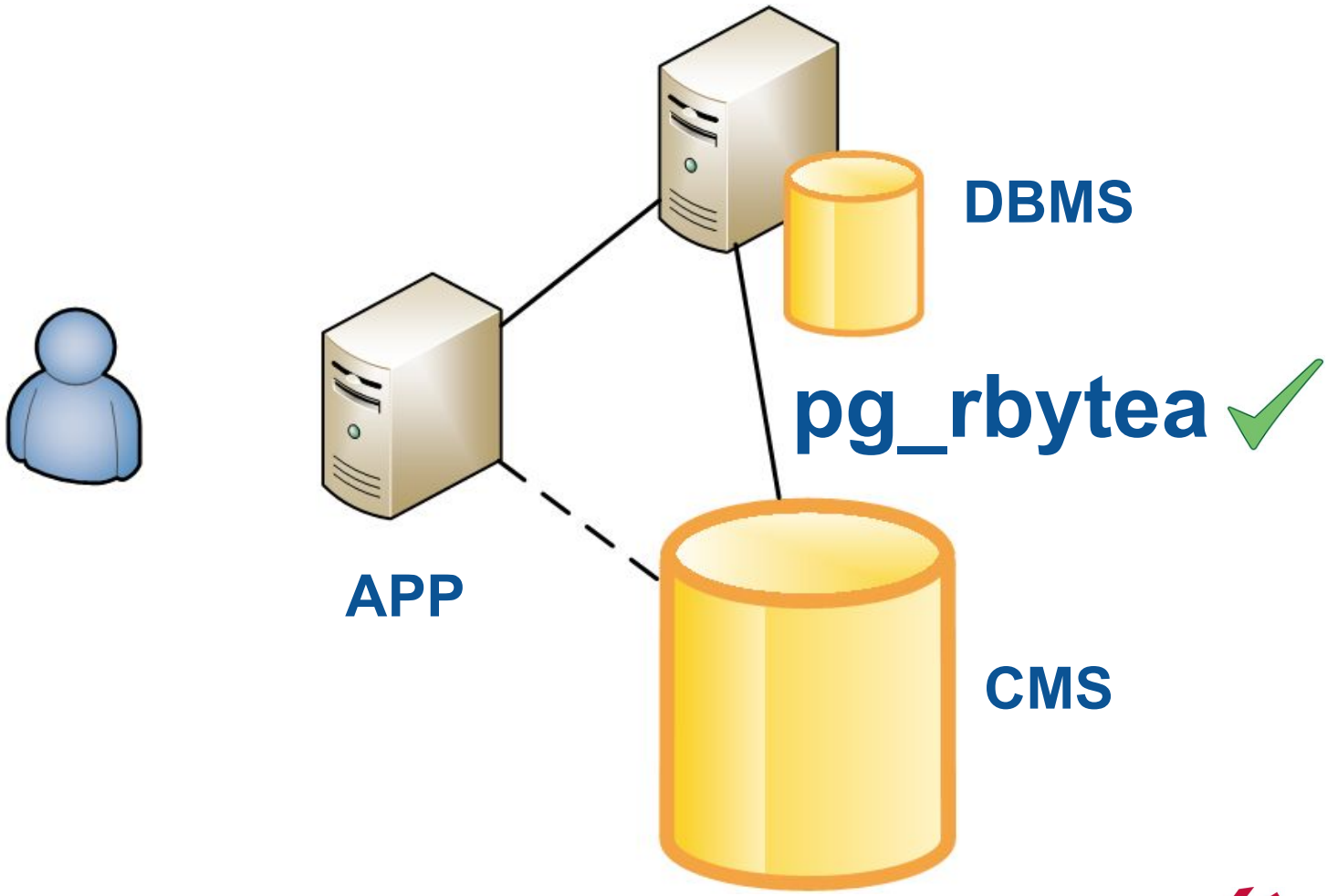


PostgreSQL

ceph

REDSYS

Предпосылки



PostgreSQL

ceph

REDSYS

EXTENSION pg_rbytea

pg_rbytea--1.0.sql

```
CREATE TYPE rbytea;
CREATE FUNCTION rbytea_in(cstring) RETURNS rbytea …;
CREATE FUNCTION rbytea_out(rbytea) RETURNS cstring …;
CREATE FUNCTION rbytea_recv(internal) RETURNS rbytea …;
CREATE FUNCTION rbytea_send(rbytea) RETURNS bytea …;

CREATE TYPE rbytea (
        INTERNALLENGTH = 20,
        INPUT = rbytea_in,
        OUTPUT = rbytea_out,
        RECEIVE = rbytea_recv,
        SEND = rbytea_send,
        STORAGE = plain
);

<…>
CREATE FUNCTION rbytea_create(i_type integer,i_flags integer,i_uuid uuid) RETURNS rbytea …;
<…>
```

PostgreSQL

EXTENSION pg_rbytea

pg_rbytea--1.0.sql

<...>

CREATE FUNCTION **rbytea_clone**(rbytea) RETURNS rbytea ...;

<...>

CREATE CAST (**rbytea** as **bytea**) with inout as implicit;
CREATE CAST (**bytea** as **rbytea**) with inout as implicit;

<...>

PostgreSQL

EXTENSION pg_rbytea

pg_rbytea--1.0.sql

<...>

CREATE ACCESS METHOD **rbhash** TYPE INDEX HANDLER **rbhashhandler**;

<...>

CREATE OPERATOR FAMILY **hash_rbytea_ops** USING **rbhash**;
CREATE OPERATOR CLASS **hash_rbytea_class** … ;

<...>

PostgreSQL

ceph

REDSYS

EXTENSION pg_rbytea

pg_rbytea.h

```
<...>

typedef struct rbytea
{
        rb_type_t    remote_storage_type;
        rb_flag_t    flags;
        rb_uuid_t    uuid;          /* object UUID in external storage or NULL */
}       rbytea;

<...>

#define STORAGE_TYPE_CEPH     (0x0001)    /* Ceph / Rados Object Storage */
#define STORAGE_TYPE_FS       (0x0002)    /* In the file system outside the database */
#define STORAGE_TYPE_TUPLE    (0x0003)    /* In another field of tuple in the same table entry */

<...>
```

PostgreSQL

EXTENSION pg_rbytea

_pg_rbytea.out

```
select rbytea_get_client_id(), rbytea_get_user_name(),
       rbytea_get_pool_name_active(), rbytea_get_pool_name_archive();

name                          | value
------------------------------+---------------------
rbytea_get_client_id          | test_client01
rbytea_get_user_name          | admin
rbytea_get_pool_name_active   | test-storage
rbytea_get_pool_name_archive | test-storage-archive
(1 row)

<...>
select id, rbytea_get_storage_type(data) type, rbytea_get_flags(data) flags,
      data, rbytea_get_uuid(data) uuid, rbytea_get_link_cnt(data) link_cnt,
      rbytea_get_xmin(data) xmin, rbytea_get_xmax(data) xmax
from test_rbytea where id=1;

name     | value
---------+-----------------------------------------------------
id       | 1
type     | 1
flags    | 1
data     | \x746573742d636f6e74656e742d7262797465613a010203
uuid     | 00efdf07-6fcb-4bb7-99db-5b28c388ea0d
link_cnt | 1
xmin     | 3601
xmax     | 0
(1 row)
```

PostgreSQL

REDSYS

EXTENSION pg_rbytea

${PGDATA}/conf.d/02-rados.conf

```
# pg_rbytea config section. CEPH/RADOS specific

# Client and user identification
pg_rbytea.rados_uniq_client_name = 'test_client01'    # Ussially, database name. Max 31 characters
pg_rbytea.rados_user_name = 'admin'

# CEPH pool names. Required at least two pools - regular and archive
# Pools must be created before using the extension

pg_rbytea.rados_active_pool_name = 'test-storage'           # Regular pool for storing objects
pg_rbytea.rados_archive_pool_name = 'test-storage-archive'  # Archive pool for deleted objects
                                                            # (recycled bin)

# Connectivity
pg_rbytea.rados_reconnect_retry = 1       # The number of attempts to restore a connection to
                                          # the storage at an unexpected break
pg_rbytea.rados_part_write_len = 4194304  # (4 MB) Block size when recording large objects

# Object lock settings
pg_rbytea.rados_lock_duration = 2000      # (in microseconds) The duration of blocking
                                          # during operations with objects
pg_rbytea.rados_lock_wait_time = 500      # The initial length of the pause while
                                          # waiting for the unlocking of the object
pg_rbytea.rados_lock_wait_attempt = 10    # The number of attempts at blocking an object
```

PostgreSQL

ceph

REDSYS

EXTENSION pg_rbytea

pg_rbytea--1.0.sql

<...>

CREATE CAST (**rbytea** as **uuid**) with function rbytea_get_uuid(rbytea) as implicit;

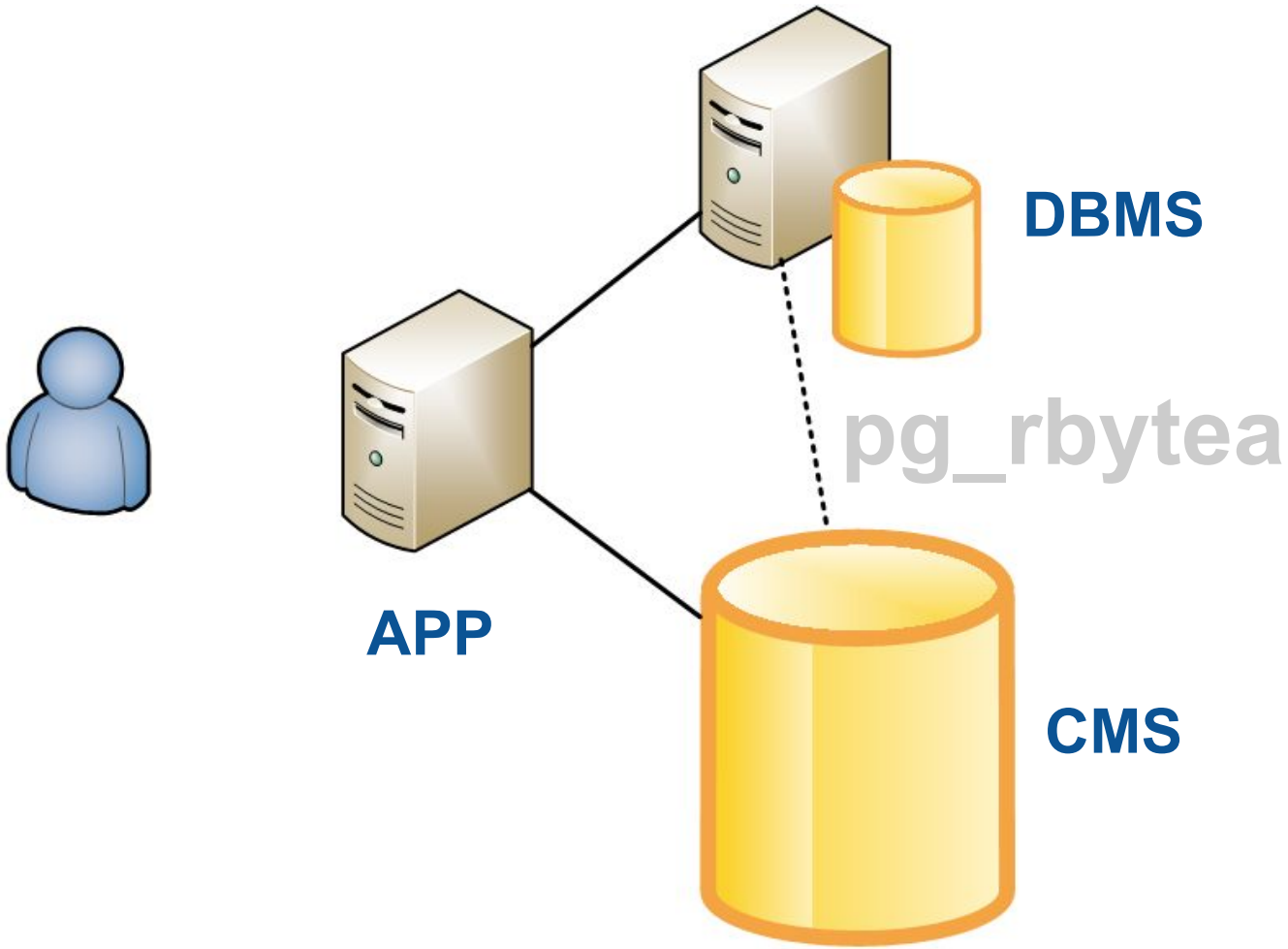
<...>

PostgreSQL

ceph

REDSYS

EXTENSION pg_rbytea



PostgreSQL

ceph

REDSYS

CEPH

THE FUTURE OF STORAGE™
Ceph is a unified, distributed storage system designed for excellent performance, reliability and scalability.

ceph.com

Ceph — отказоустойчивое распределенное хранилище данных, работающее по протоколу TCP.
Одно из базовых свойств Ceph — масштабируемость до петабайтных размеров.

Ceph предоставляет на выбор три различных абстракции для работы с хранилищем:

- абстракцию объектного хранилища (RADOS Gateway),
- блочного устройства (RADOS Block Device) или
- POSIX-совместимой файловой системы (CephFS).

habrahabr.ru

PostgreSQL

ceph

REDSYS

Ограничения и риски

- Типы хранения
  - Ceph - реализован
  - В локальной файловой системе - в разработке
  - В той же записи - в разработке
- Применимость по версиям, PostgreSQL:
  - 9.6 - разработан, наиболее оттестирован
  - 10 - в процессе тестирования
- Ceph / Rados:
  - librados версий от 0.94 (hummer) до 11.2.1 (kraken)
  - неиспользование palloc / pfree
- Опыт миграции
  - Имеется только для миграции в Ceph, одномоментно
  - Vacuum для Ceph реализован не полностью

PostgreSQL

ceph

REDSYS

PGConf.Russia 2018

Российская конференция по PostgreSQL

Валерий Косарев
valeriy.kosarev@redsys.ru
valeriy.kosarev@gmail.com

Компания REDSYS
www.redsys.ru

Sources:
https://github.com/val5244/pg_rbytea

Спасибо за внимание.

PostgreSQL