
OTUS
ОНЛАЙН-ОБРАЗОВАНИЕ
Онлайн-образование

Меня хорошо видно && слышно?

Ставьте + , если все хорошо

Напишите в чат, если есть проблемы

заодно проверяем, включена ли запись занятия

Включил Юджин запись ли ты

Работа с кластером высокой доступности для PostgreSQL

Аристов Евгений

telegram @AEugene

https://aristov.tech

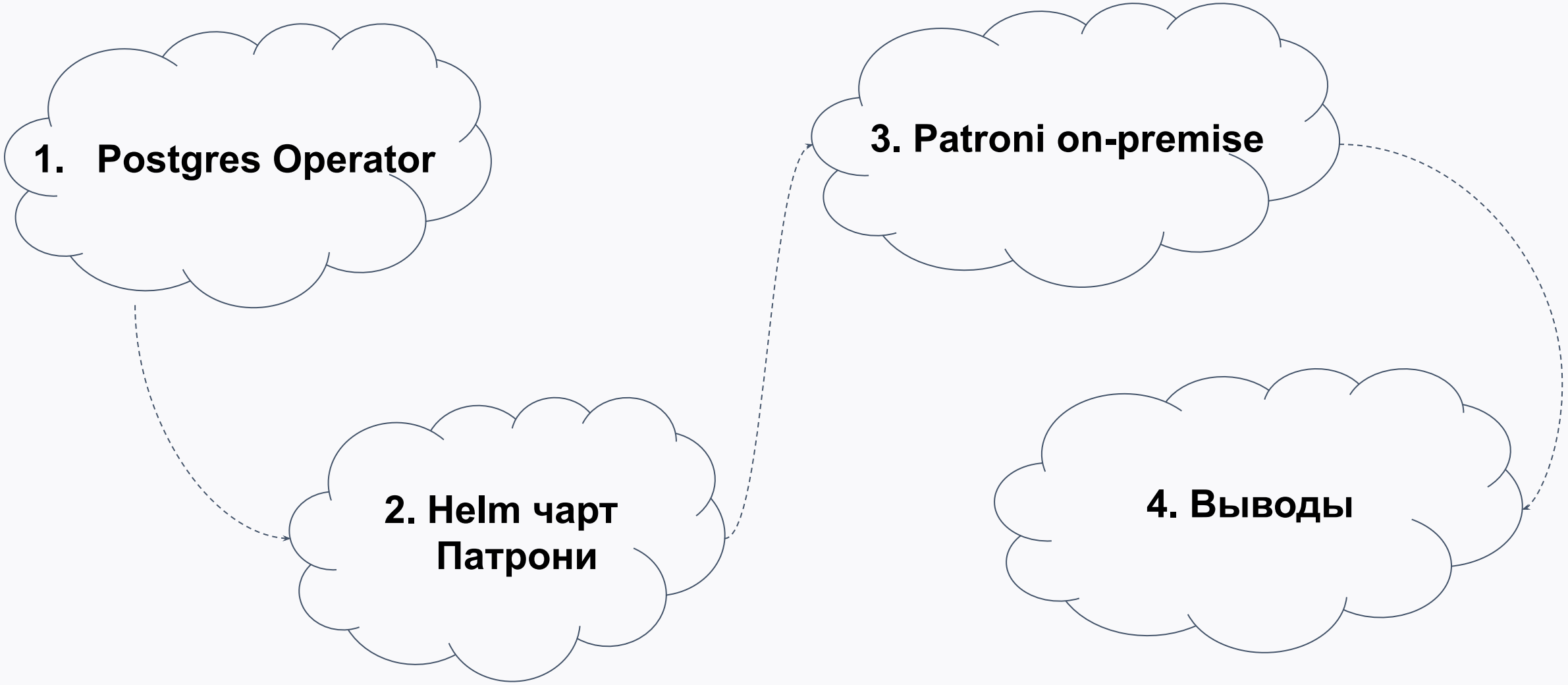
Правила вебинара

Активно участвуем

Задаем вопрос в чат

Вопросы вижу в чате, могу ответить не сразу

Маршрут вебинара



Цели вебинара | После занятия вы сможете

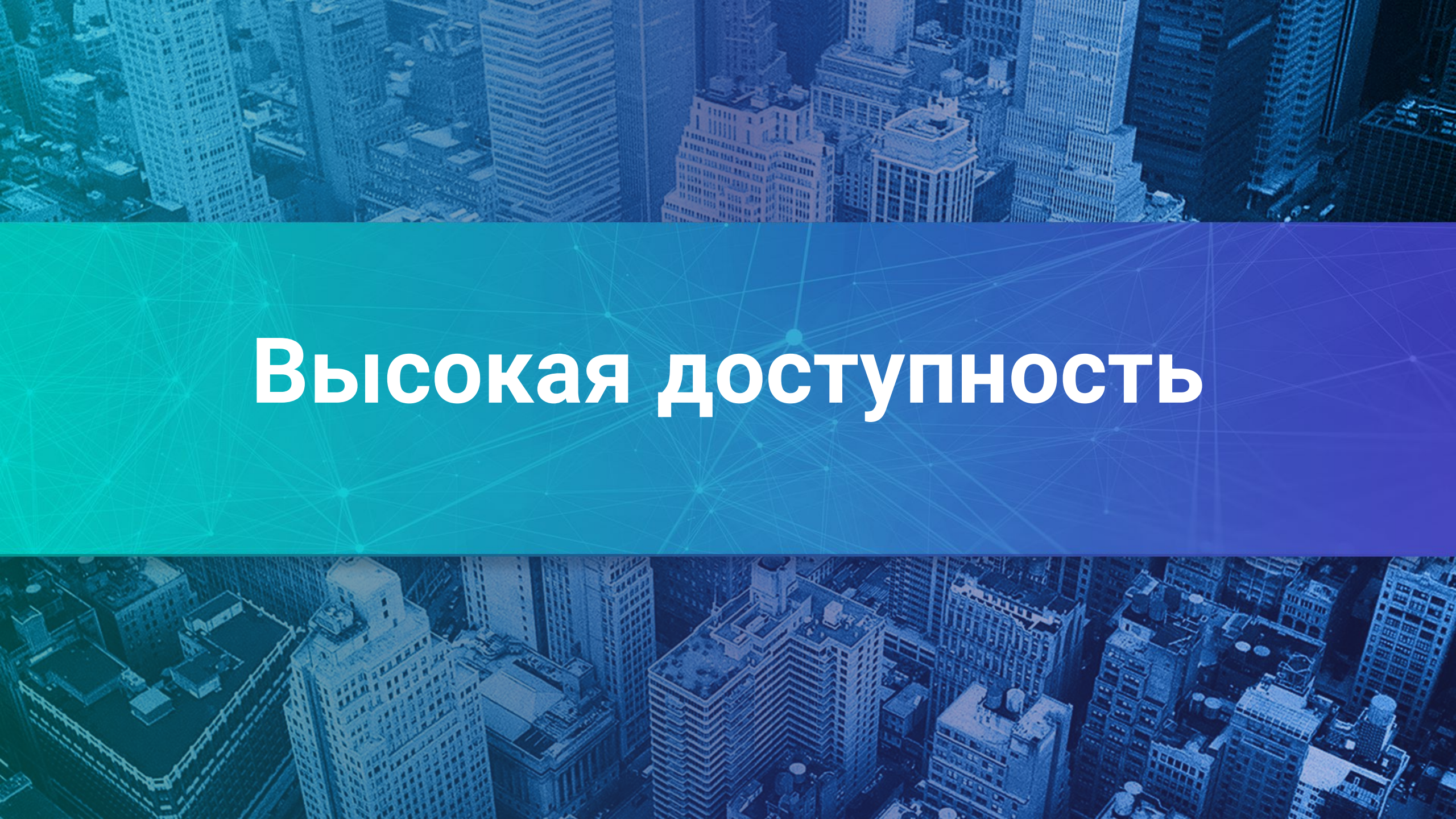
1 Уметь настраивать кластер Patroni через чарт helm

2 Уметь настраивать классический кластер Patroni

Смысл | Зачем вам это уметь, в результате:

1 Выбрать оптимальный вариант высокой доступности для PostgreSQL

2 Уметь настроить высокодоступный кластер Patroni своими руками

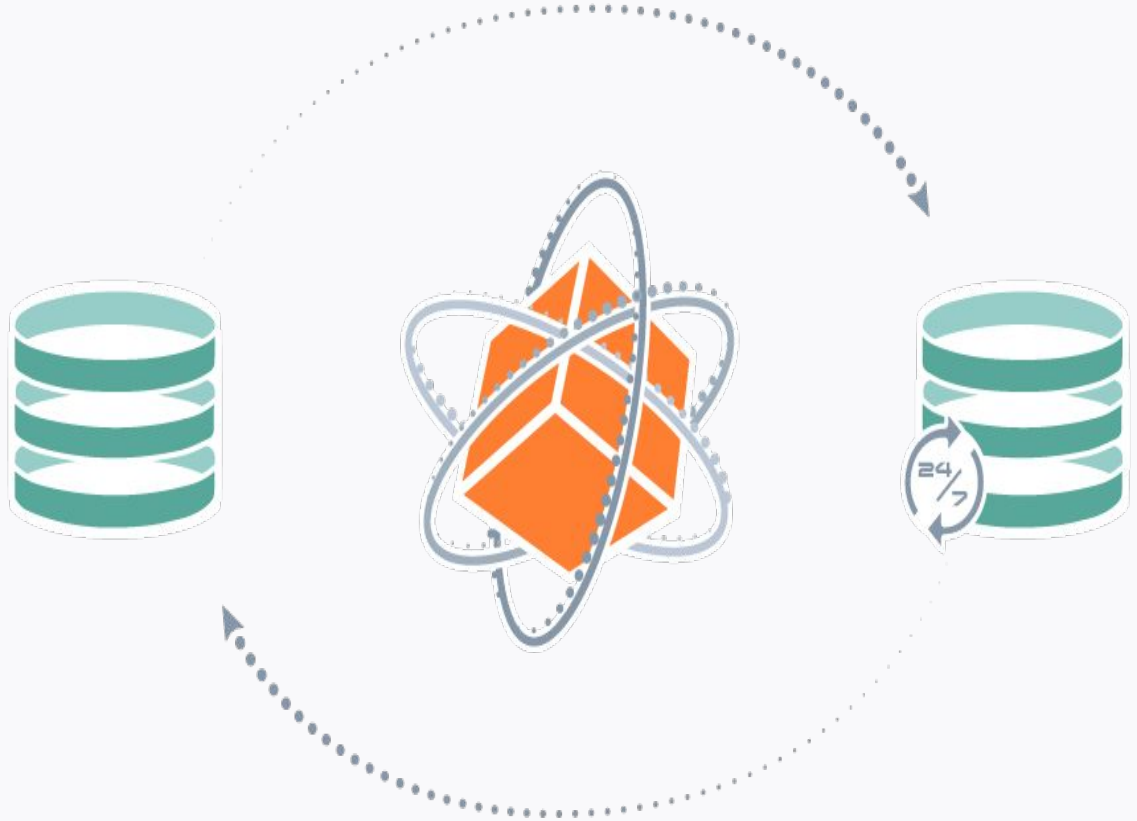


Высокая доступность

Высокая доступность

- High Availability, или HA
- измеряется в 9-ках
- все хотят 99,999 - 5 минут простоя в год
- все предлагают 99,99 - час простоя в год
- метод измерения и критерий доступности - тот еще вопрос
- например работает но жутко тормозит - это как?

[Калькулятор SLA: 99.9% аптайм]



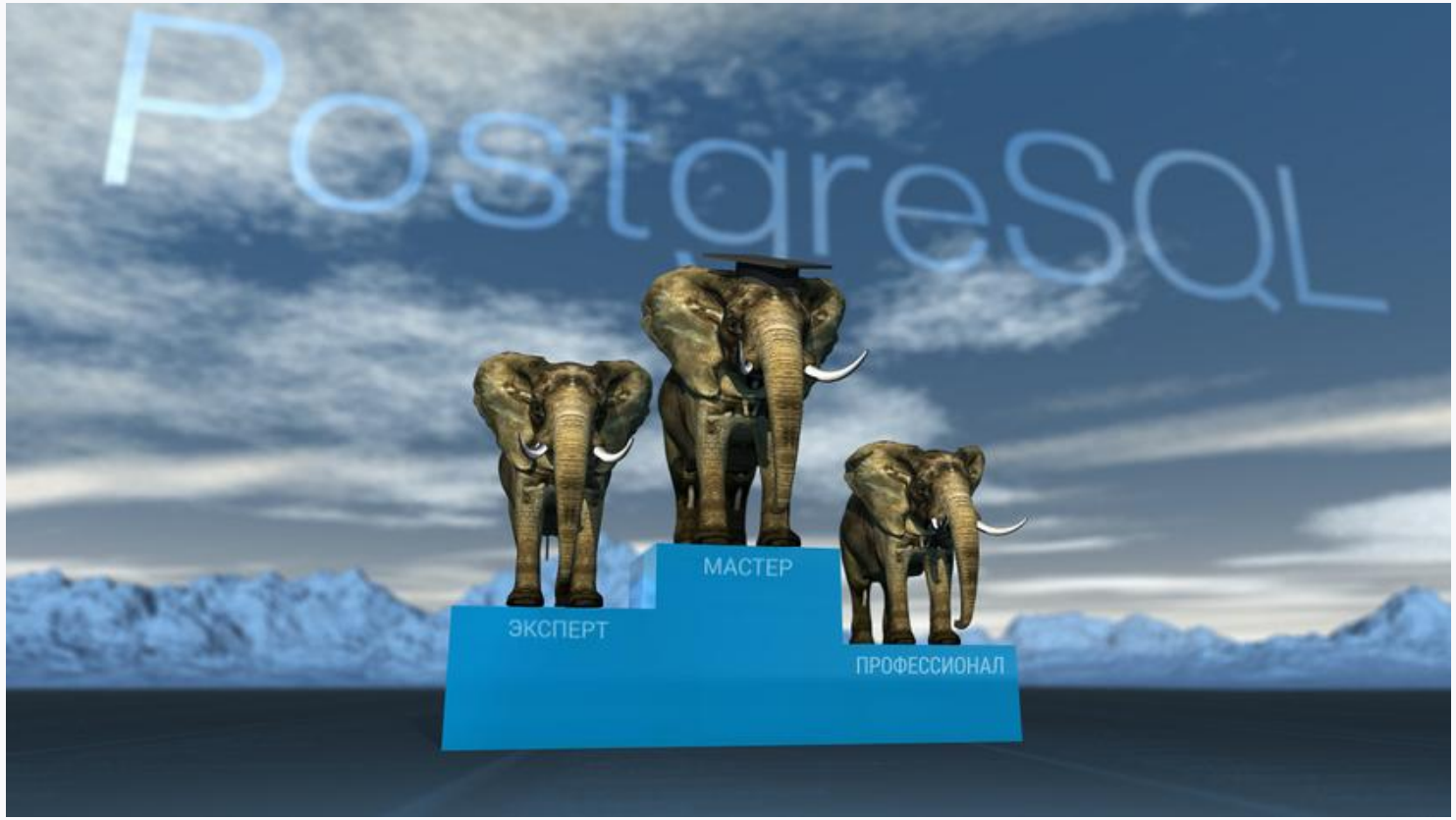
HA кластеры PostgreSQL

классические

- Pgpool II
- pg_bouncer + haproxy
- pg-auto-failover
- repmgr

cloud native

- **patroni**
- stolon
- slony
- ClusterControl



Patroni

Patroni is a template for you to create your own customized, high-availability solution using Python and - for maximum accessibility - a distributed configuration store like ZooKeeper, etcd, Consul or Kubernetes.

patroni

Теория

Patroni

Обсудим основные проблемы

https://habr.com/ru/post/504044/

Patroni

Функции DCS (*distributed* control system)

- etcd (или Consul, Zookeeper) хранят информацию о том, кто сейчас лидер
- DCS хранит конфигурацию кластера
- помогает решить проблему с партиционированием сети
- STONITH
- Неплохо бы иметь watchdog (Например, Nomad by HashiCorp)

Patroni

Consul

- Service check
- + Consul templates
- Есть GUI =)
- Есть свой DNS
- Patroni может анонсировать master/replica
- ETCD при большой загрузке замечен в высокой нагрузке на дисковую подсистему

с версии 2.0 нативно поддерживает выборы нового мастера по рафт протоколу без использования etcd/consul

https://raft.github.io/

Patroni

Зачем нужен Patroni

- PostgreSQL не умеет взаимодействовать с etcd
- Демон на питоне будет запущен рядом с PostgreSQL
- Демон умеет взаимодействовать с etcd
- Демон принимает решение promotion/demotion

Patroni

Создать свой первый кластер Patroni:

patroni /etc/patroni.yml OR **systemctl start patroni.service**

INFO: Selected new etcd server http://10.128.0.48:2379

INFO: Lock owner: None; I am pg01

trying to bootstrap a new cluster§

LOG: listening on IPv4 address "10.128.0.49", port 5432

INFO: establishing a new patroni connection to the postgres cluster

INFO: Lock owner: pg01; I am pg01

INFO: no action. i am the leader with the lock

Patroni

Состояние кластера

- patronictl - утилита для управления кластером
- patronictl -c /etc/patroni.yml list

```
[root@pg01 ~]# patronictl -c /etc/patroni.yml list
+----------+--------+-------------+--------+---------+----+-----------+
| Cluster  | Member |     Host    |  Role  |  State  | TL | Lag in MB |
+----------+--------+-------------+--------+---------+----+-----------+
| postgres |  pg01  | 10.128.0.47 |        | running |  7 |       0.0 |
| postgres |  pg02  | 10.128.0.46 | Leader | running |  7 |       0.0 |
| postgres |  pg03  | 10.128.0.45 |        | running |  7 |       0.0 |
+----------+--------+-------------+--------+---------+----+-----------+
```

Patroni

Автоматический Failover

systemctl stop patroni - любой другой способ протестировать failover =)

• 30 секунд по умолчанию на истечение ключа в DCS

• После чего Patroni стучится на каждую ноду в кластере и спрашивает, не мастер ли ты, проверяет WAL логи, насколько близки они к мастеру. В итоге если WAL логи у всех одинаковые то, промоутится следующий по порядку

• Опрос нод идёт параллельно

Patroni

Важные параметры

Обновление данных в DCS идет циклично:

- loop_wait - промежуток в секундах между попытками обновить ключ лидера.
- retry-timeout - через сколько будем ретраить попытку обновить ключ
- ttl - время жизни ключа лидера. Рекомендация: как минимум loop_wait + retry_timeout, но вообще таким комфортным, чтобы избежать нескольких медленных/неудавшихся вызовов к DCS
- maximun_lag_on_failover - максимальное отставание ноды от лидера для того, чтобы участвовать в выборах
- synchronous_mode - вкл/выкл синхронной реплики
- synchronous_mode_strict - вкл/выкл строго синхронного режима, чтобы мастер останавливался при смерти синхронной реплики

https://patroni.readthedocs.io/en/latest/replication_modes.html

Patroni

Редактирование конфигурации

patronictl -c /etc/patroni.yml edit-config

Ручной Switchover:

patronictl -c /etc/patroni.yml switchover

Перезагрузка

- patronictl -c /etc/patroni.yml restart postgres pg02
- Применение новых параметров требующих обязательной перезагрузки

Реинициализация

- patronictl -c /etc/patroni.yml reinit postgres pg03
- Реинициализирует ноду в кластере. Т.е. по сути удаляет дата директорию и делает pg_basebackup, если это поведение не изменено параметром create_replica_method

Patroni

Локальная конфигурация

Что делать если нужно поменять конфигурацию PostgreSQL только локально:

- patroni.yml
- postgresql.base.conf
- ALTER SYSTEM SET - имеет наивысший приоритет

Некоторые параметры, такие как: max_connections, max_locks_per_transaction, wal_level, max_wal_senders, max_prepared_transactions, max_replication_slots, max_worker_processes **не могут быть переопределены локально - Patroni их перезаписывает**.

https://patroni.readthedocs.io/en/latest/SETTINGS.html?highlight=custom_conf#postgresql

Patroni

Monitoring

Проверка запущен ли PostgreSQL мастер:

• GET /master - должно возвращать 200 ТОЛЬКО для одной ноды, остальные 503 и наоборот GET /replica

Проверка работают ли реплики

• GET /patroni с мастера должно возвращать replication:[{state: streaming}]

для всех реплик

Запущен ли сам PostgreSQL:

• GET /patroni должен возвращать state:running для каждой ноды

Отставание реплики:

• GET /patroni - xlog: location с реплик не должен быть далеко от этого же

параметра на мастере

Patroni

Роутинг трафика

- HAProxy, TCP Proxy (NGINX)
- Pgbouncer (pgPool, Odyssey)

Patroni

Пользовательские скрипты. ХУКИ!

postgresql:

callbacks:

    on_start: /opt/pgsql/pg_start.sh

    on_stop: /opt/pgsql/pg_stop.sh

    on_role_change: /opt/pgsql/pg_role_change.sh

Patroni

Tags

- nofailover (true/false) - в положении true нода никогда не станет мастером
- noloadbalance (true/false) - /replica всегда возвращает код 503
- clonefrom (true/false) - patronictl выберет предпочтительную ноду для pgbasebackup
- nosync (true/false) - нода никогда не станет синхронной репликой
- replicatefrom (node name) - указать реплику с которой снимать реплику

Patroni

Switchover vs failover

- Switchover
    - Переключение роли Мастера на новую ноду. Делается вручную, по сути плановые работы
- Failover
    - Экстренное переключение Мастера на новую ноду
    - Происходит автоматически
    - Ручной вариант - manual failover - только когда не система не может решить на кого переключать

Patroni

Режим паузы

- Отключается автоматический failover
- Ставиться глобальная пауза на все ноды
- Проведение плановых работ, например с etcd или обновление PostgreSQL

Тем не менее:

- Можно создавать реплики
- Ручной switchover возможен
- patronictl -c /etc/patroni.yml pause|resume

Patroni

Создание реплики из бекапа

• По умолчанию реплика создается с помощью утилиты pg_basebackup

• Это поведение можно переопределить параметром create_replica_methods

• Важно, обязательно нужно указать basebackup, иначе если из бекапа не получится, то реплика не заведется.

```
postgresql:
    create_replica_methods:
        - probackup
        - basebackup
probackup:
    command: "ssh dbbackup@10.23.2.163 ‘bash /var/backup/pg_restore.sh'"
    no_params: True
basebackup:
    max-rate: '100M'
```

Patroni

Истории аварий с Patroni, или Как уронить PostgreSQL-кластер

Patroni

Параметры в Патрони:

https://patroni.readthedocs.io/en/latest/SETTINGS.html?highlight=custom_conf#postgresql

Практика

Postgres operator

Helm чарт

кубер в GKE

на базе патрони

https://github.com/zalando/postgres-operator

https://github.com/zalando/postgres-operator/blob/master/docs/quickstart.md

https://postgres-operator.readthedocs.io/en/latest/

https://severalnines.com/database-blog/how-achieve-postgresql-high-availability-pgbouncer

https://www.pgbouncer.org/config.html

https://www.pgbouncer.org/faq.html

Postgres operator

Helm чарт

кубер в GKE

новый от Crunchy data

https://www.postgresql.org/about/news/pgo-the-crunchy-postgres-operator-v5-released-fully-declarative-postgres-2258/

Patroni

Helm чарт

кубер в GKE

https://patroni.readthedocs.io/en/latest/

https://github.com/helm/charts/tree/master/incubator/patroni

https://patroni.readthedocs.io/en/latest/kubernetes.html

walE.enable

используем CDS кубика, на выбор etcd, consul, zookeeper

из минусов:

- нет сервиса нормального
- нет лоад балансера
- и прокси

Patroni on-premise

Полуручное раскатывание

https://github.com/zalando/patroni

Patroni on-premise

Ручное раскатывание

Построение кластера PostgreSQL высокой доступности с использованием Patroni, etcd, HAProxy

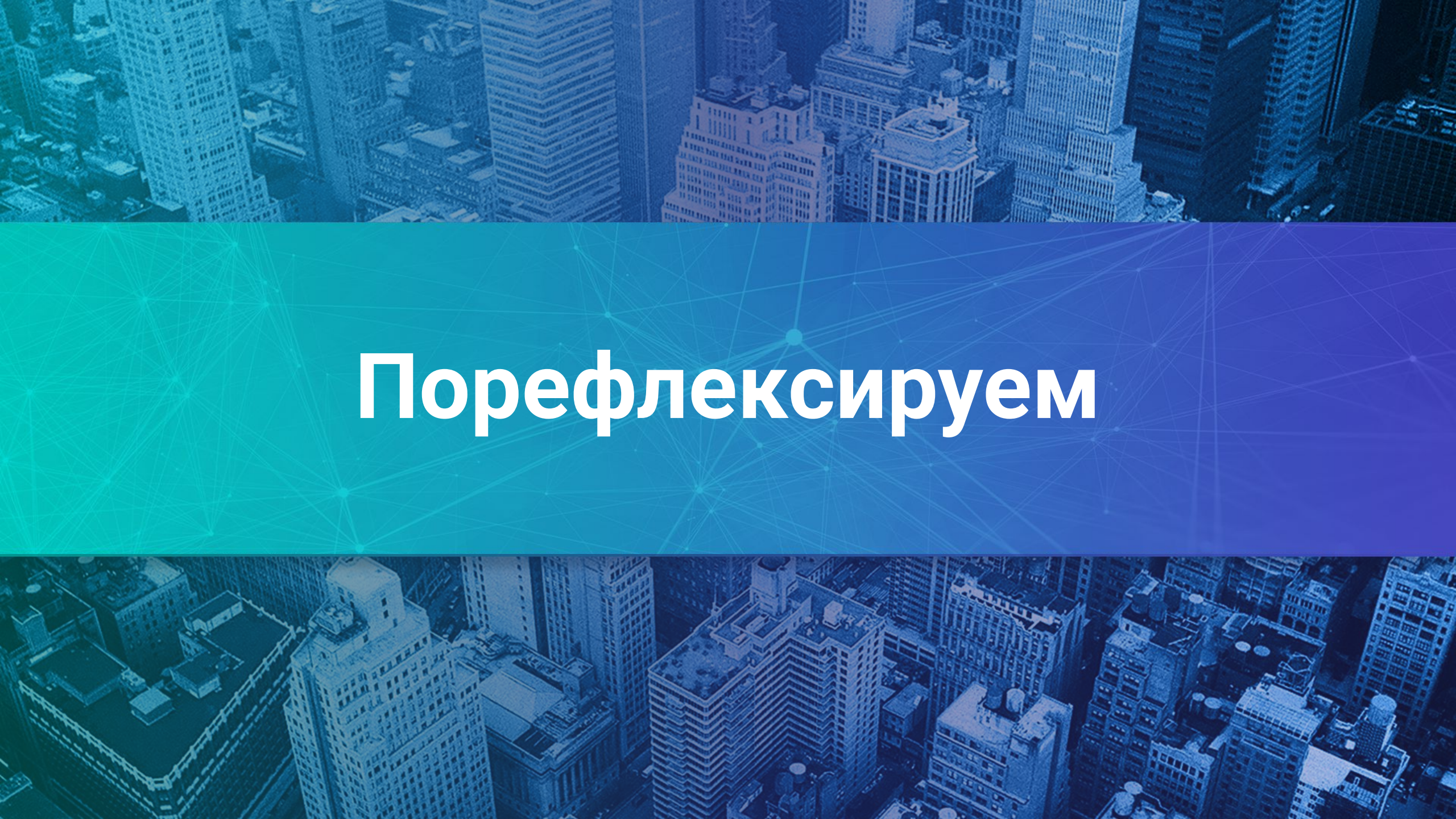
Заряжай Patroni. Тестируем Patroni + Zookeeper кластер (Часть первая) / Блог компании VS Robotics / Хабр

Заряжай Patroni. Тестируем Patroni + Zookeeper кластер (Часть вторая) / Блог компании VS Robotics / Хабр

минитест

https://forms.gle/HZbRecH6sgT9qtb67

2-3 минуты


Порефлексируем

Вопросы?

- **Какой HA кластер понравился?**

ДЗ

ДЗ

Вариант 1

How to Deploy PostgreSQL for High Availability

Вариант 2

Introducing pg_auto_failover: Open source extension for automated failover and high-availability in PostgreSQL

Для гурманов

Настройка Active/Passive PostgreSQL Cluster с использованием Pacemaker, Corosync, и DRBD (CentOS 5,5)

ДЗ

**** Создать два кластера GKE в разных регионах***

Установить на первом Patroni HA кластер

Установить на втором Patroni Standby кластер

Настроить TCP LB между регионами

Сделать в каждом регионе по клиентской ВМ

Проверить как ходит трафик с клиентской ВМ

Описать что и как делали и с какими проблемами столкнулись


Заполните, пожалуйста,
опрос о занятии по ссылке в чате
https://otus.ru/polls/43429/

Спасибо за внимание!
Приходите на следующие вебинары

Аристов Евгений