

Реляционные базы данных: Базы данных в облаке



Нарек
Татевосян



Нарек Татевосян

Enterprise Architect Team Lead, Yandex.Cloud





Вспоминаем прошлые занятия

Вопрос: Какие бывают типы репликации баз данных и чем они отличаются ?

Вспоминаем прошлые занятия

Вопрос: Какие бывают типы репликации баз данных и чем они отличаются ?

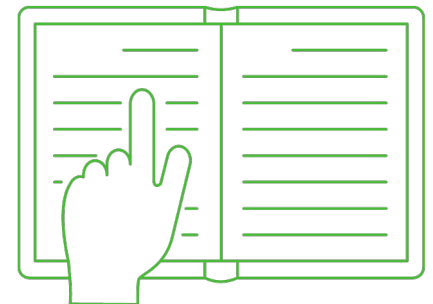
Ответ:

- **Синхронная.** Если данная реплика обновляется, все другие реплики того же фрагмента данных также должны быть обновлены в одной и той же транзакции. Логически это означает, что существует лишь одна версия данных.
- **Асинхронная.** Обновление одной реплики распространяется на другие спустя некоторое время, а не в той же транзакции. Таким образом, при асинхронной репликации вводится задержка, или время ожидания, в течение которого отдельные реплики могут быть фактически неидентичными.

Предисловие

На этом занятии мы:

- поговорим о том, как работают управляемые базы данных и чем такая база данных отличается от своей;
- познакомимся с сервисами Managed Database for Postgresql.





План занятия

1. [Как работает кластер?](#)
2. [Как подключиться к кластеру?](#)
3. [Как происходит восстановление после отказа?](#)
4. [Эксплуатация кластера](#)
5. [Резервное копирование в кластере](#)
6. [Доступность и SLA](#)
7. [Квоты и лимиты](#)
8. [Как переехать в управляемую базу?](#)
9. [Итоги](#)
10. [Домашнее задание](#)



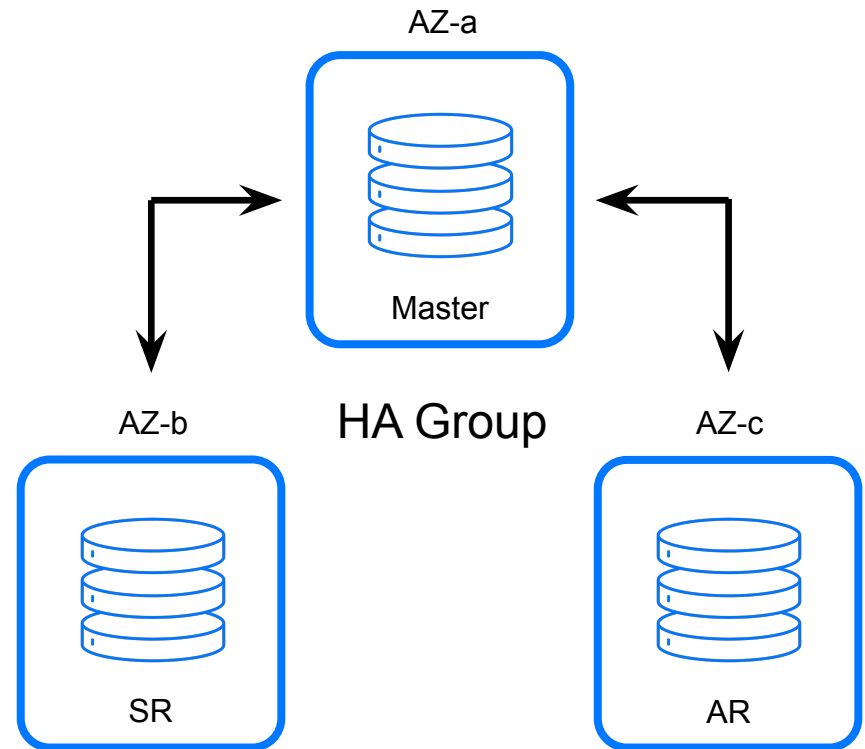
Как работает кластер?

Что такое Managed Databases?

1. Возможность развернуть базу нажатием кнопки.
2. Можно менять конфигурацию.
3. Доступны кросс-ДЦ кластеры.

Виды конфигураций:

- Одноузловая (по умолчанию);
- Кластер (минимум 2 узла).



В каких инструментах доступен сервис?

- Web Console
- YC CLI tool
- Service Provider for Terraform
- REST API
- Python / GO / Java SDK

Разделение ролей

Команда MDB Yandex Cloud

- Доступность и отказоустойчивость
- Резервное копирование
- Консоль мониторинга
- Обновления (как минорные, так и мажорные)
- Поддерживаемые клиенты
- Техническая поддержка.

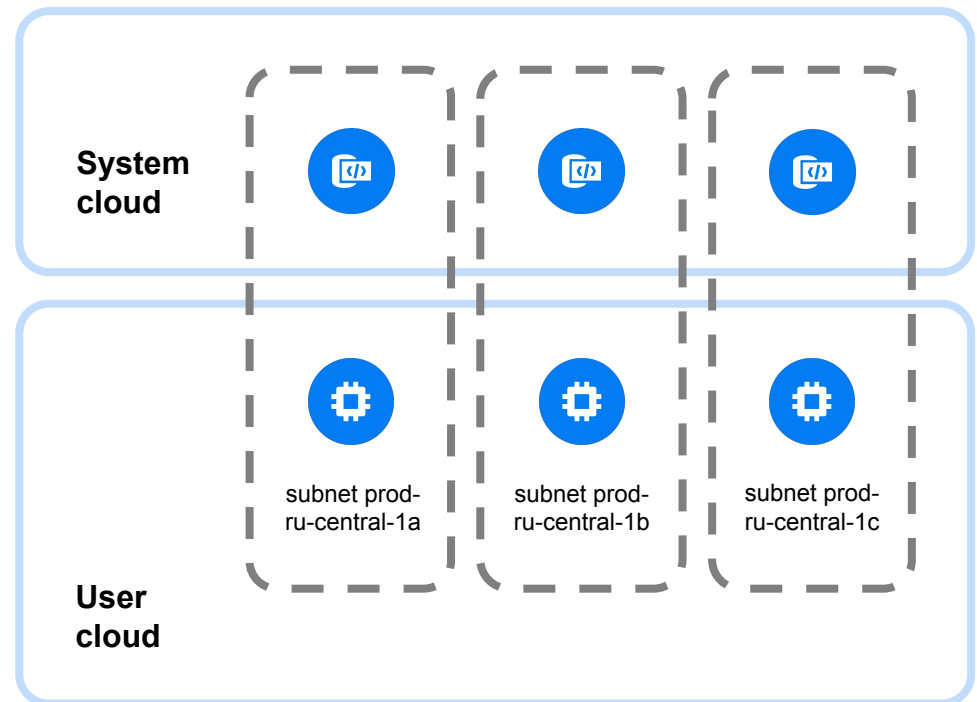
Пользователь

- Схема данных
- Запросы
- Мониторинг производительности
- Контроль занимаемого места

Как выглядит кластер с точки зрения Облака?

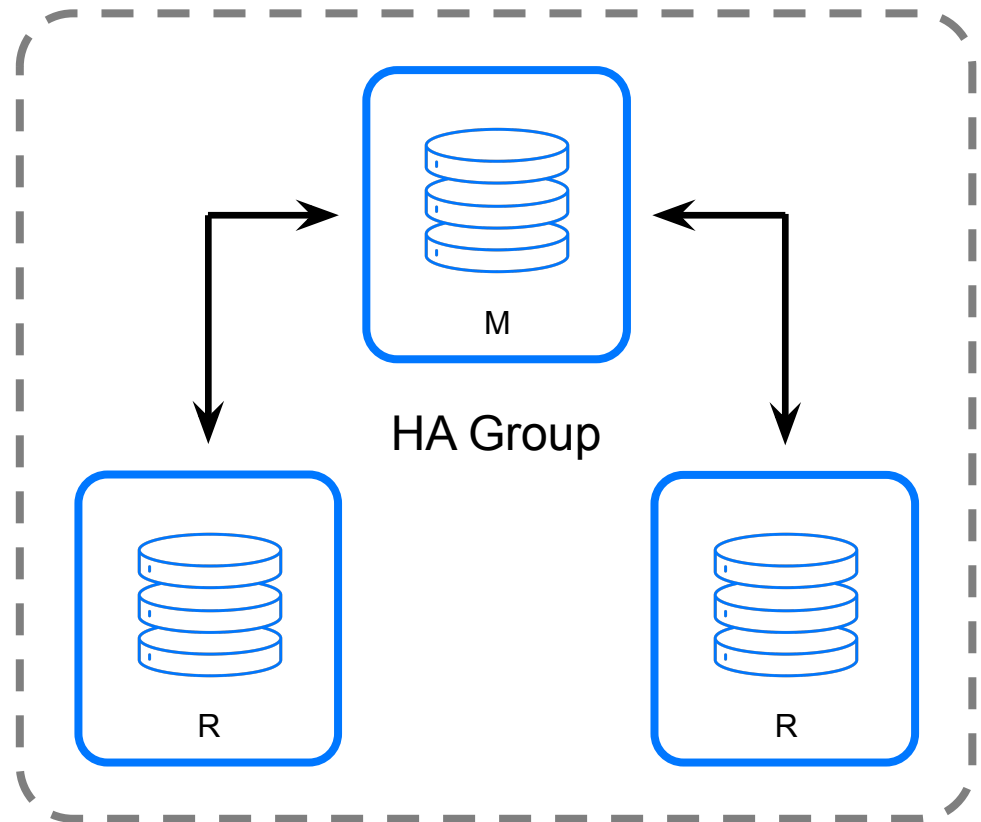
Хосты кластера — это VM сервиса Yandex Compute Cloud, подключенные в вашу сеть.

- **Вам доступны:** базы данных для вашего приложения, ваши данные, настройки базы и кластера
- **Вам недоступны:** виртуальные машины сервиса, системные пользователи и базы.



Как выглядит кластер с точки зрения СУБД?

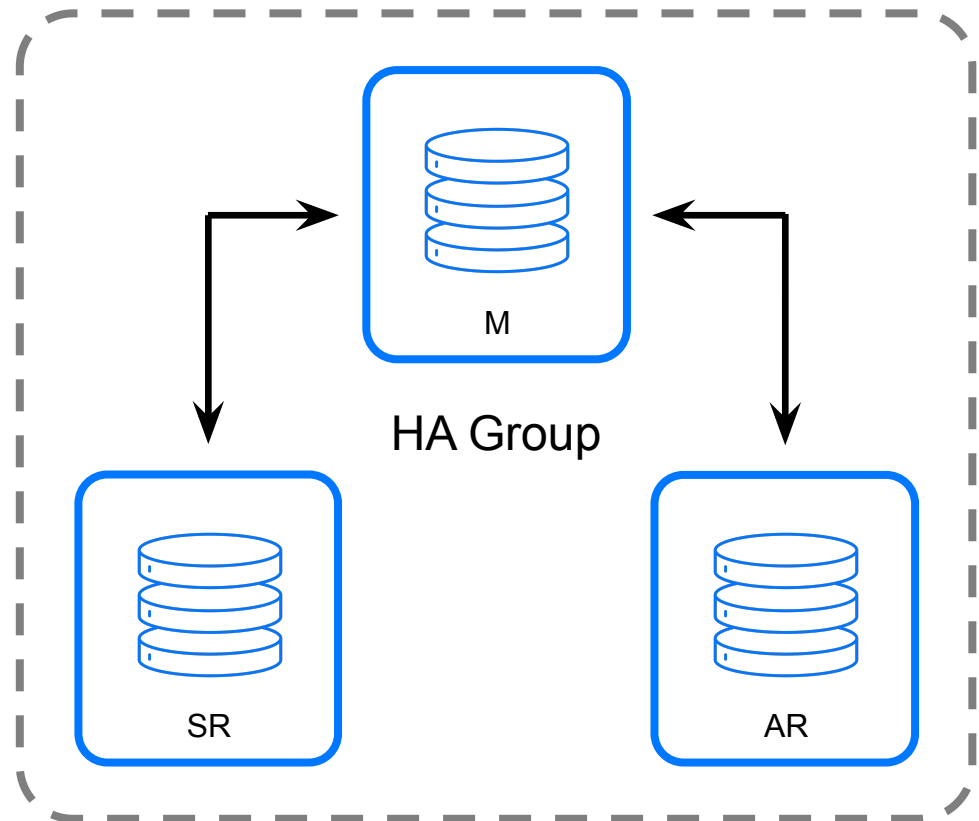
- Single Master кластер
- Хосты собираются в HA-группу
- Любой из хостов в HA-группе может стать мастером
- Все реплики всегда синхронизируются с мастера



Репликация. Часть 1 (синхронная)

- Для подтверждения записи нужен коммит от синхронной реплики
- Первая реплика синхронная, остальные асинхронные

Доступно в Yandex Managed Service for PostgreSQL

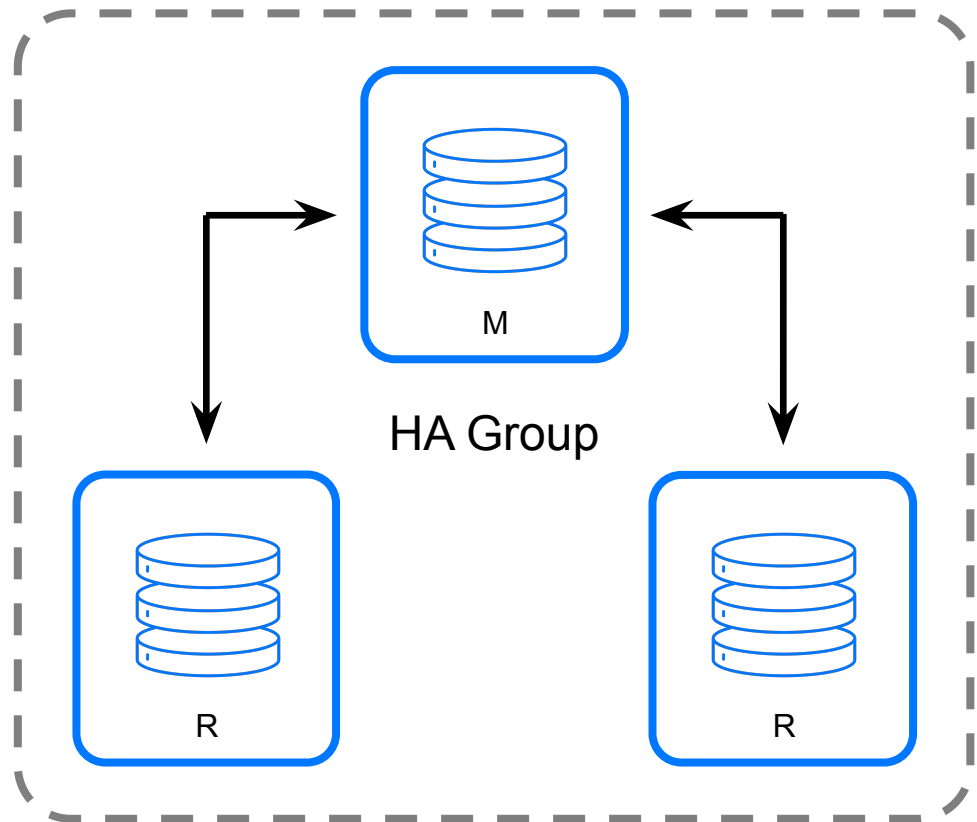


Репликация. Часть 2 (полусинхронная)

- Для подтверждения записи нужен коммит от кворума хостов
- Все реплики полусинхронные

Доступно в Yandex Managed Service for MySQL.

Планируется в Yandex Managed Service for PostgreSQL.

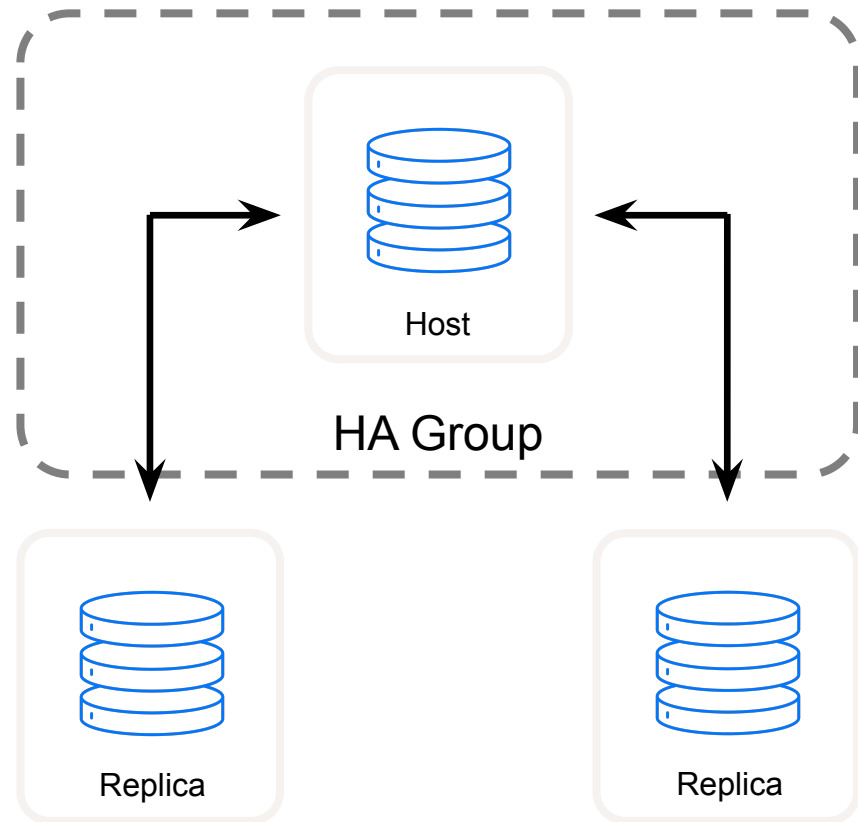



Репликация. Часть 3 (каскадная)

- Асинхронная реплика с одного из хостов HA-группы
- Доступна только на чтение, недоступна на запись
- Не является частью HA-группы

Доступно в Yandex Managed Service for PostgreSQL.

Планируется в Yandex Managed Service for MySQL.

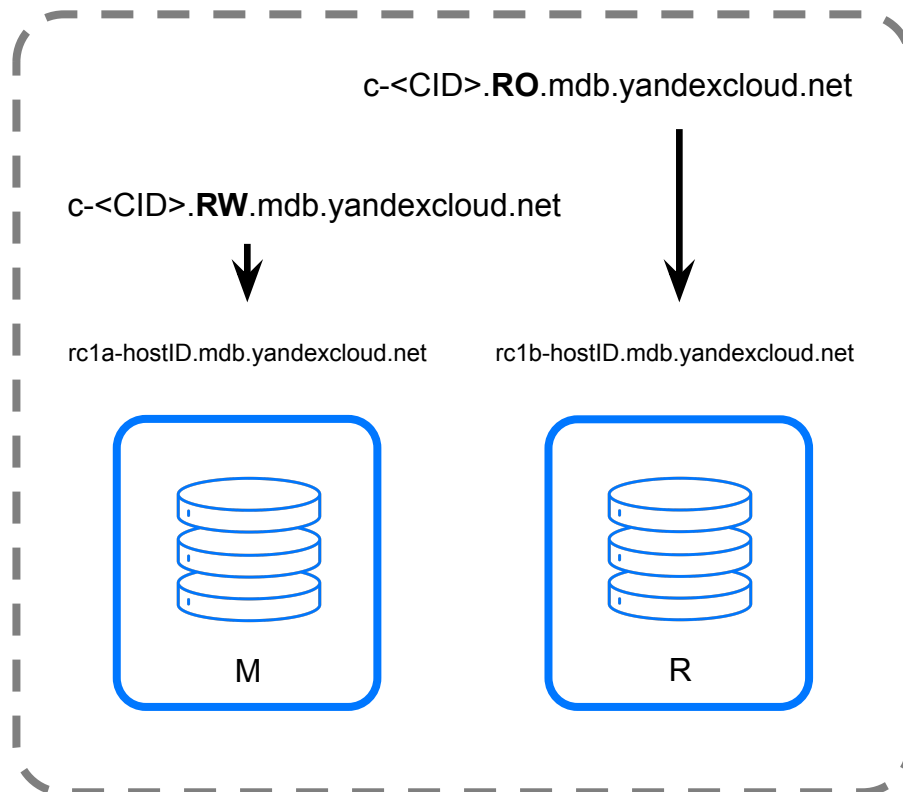




Как подключаться к кластеру?

По доменному имени

- Подключение к кластеру использует DNS
- У каждого хоста есть неизменяемый FQDN
- У каждого кластера есть C-NAME запись, ведущая на мастер
- У каждого кластера есть C-NAME запись, ведущая на одну из реплик

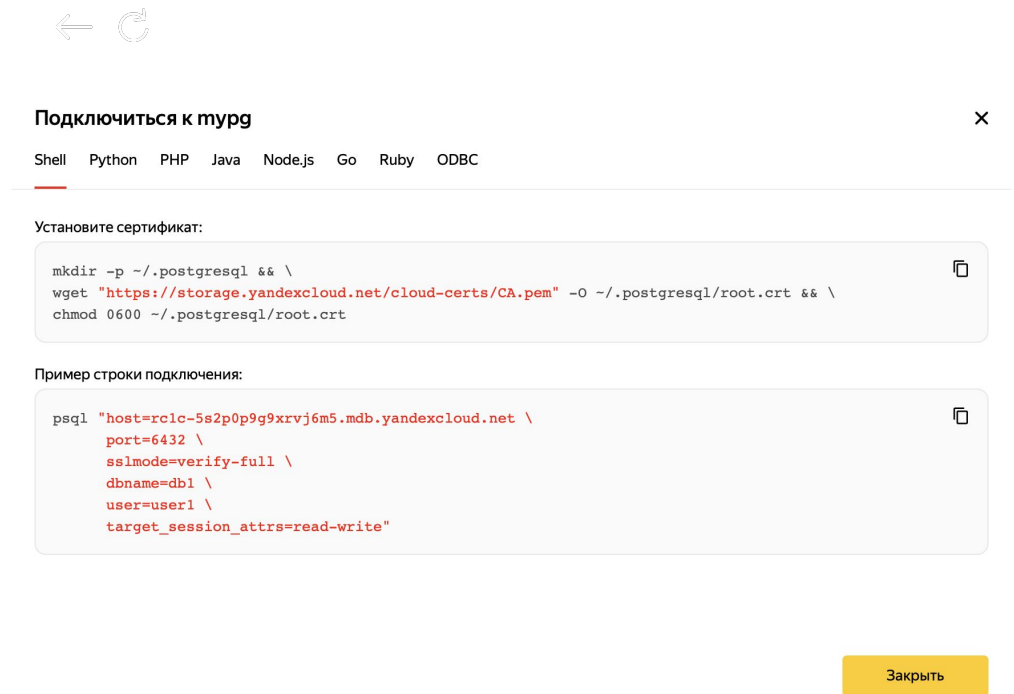


В Yandex Managed Service

for PostgreSQL вы подключаетесь в pooler Odyssey, в Yandex Managed Service for MySQL — в СУБД.

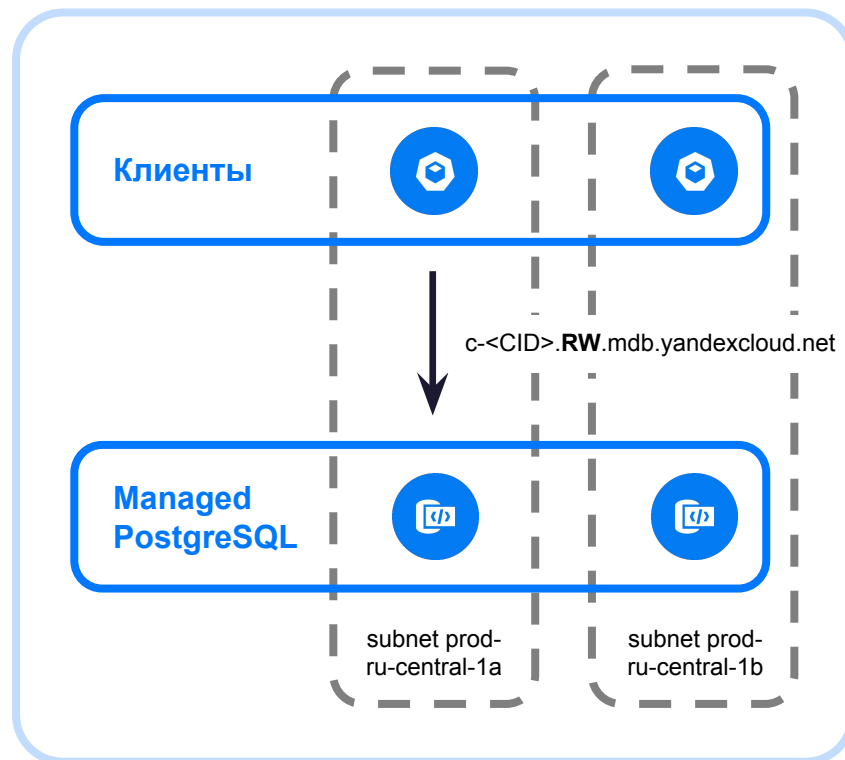
Автоматический выбор мастера в клиенте

- Если позволяет клиентская библиотека, то выбор мастера рекомендуется делать на ней, т.к. скорость восстановления после отказа будет выше, чем при использовании C-NAME
- Сейчас это доступно для PostgreSQL (libpq, JDBC)
- В MySQL доступно JDBC



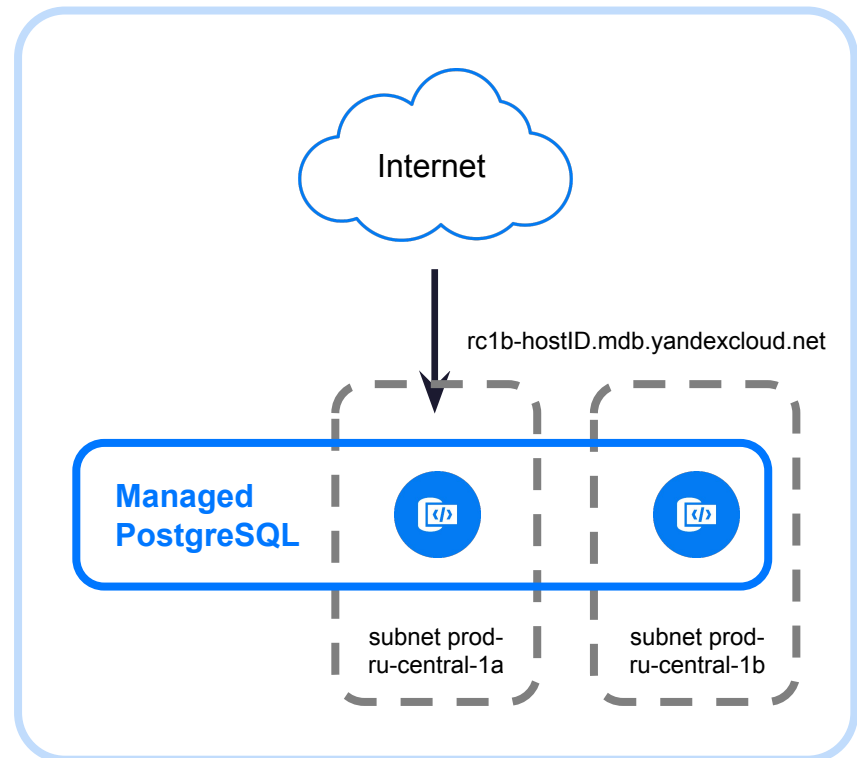
Подключение внутри сети

- DNS-записи кластера резолвятся в сети в приватные IP-адреса сети
- Такие записи доступны сервисам, живущим в сетях клиентов — например, Yandex Compute Cloud, Yandex Managed Service for Kubernetes
- Из другой сети эти DNS-записи недоступны
- Yandex DataLens может подключиться к приватному кластеру, если явно разрешить
- Для PostgreSQL доступен WebSQL через UI



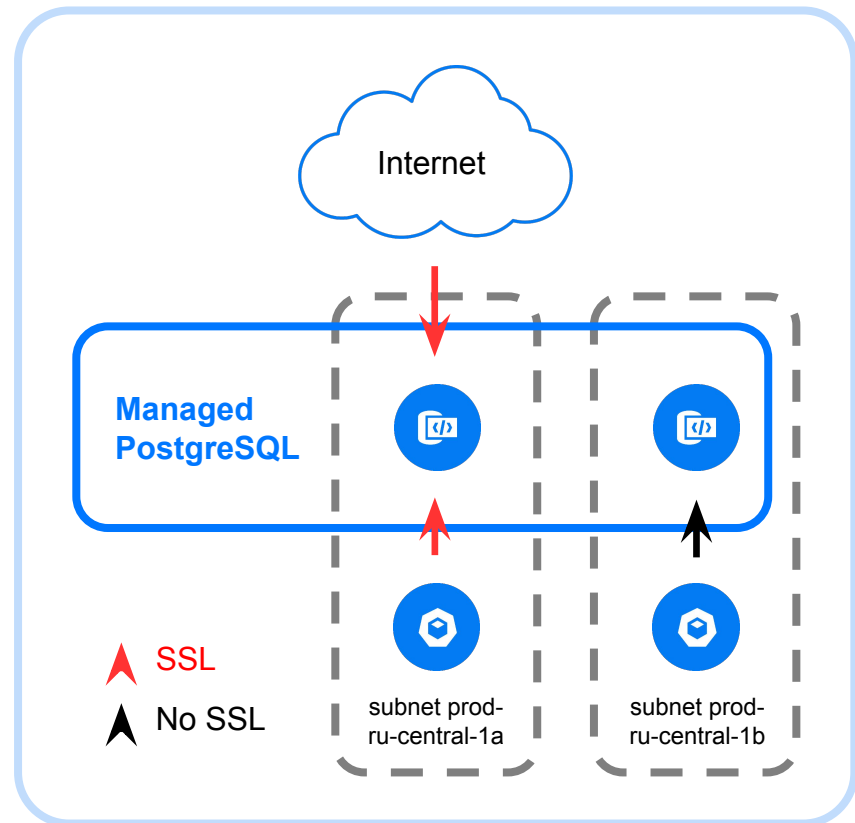
Подключение из интернета

- Можно выставить часть хостов в интернет, тогда их hostname начнут резолвиться в интернет
- Специальные C-NAME начнут резолвиться в публичные хосты
- RW C-NAME не будет резолвиться в интернет, если мастер находится на публичном хосте



Включение и отключение SSL

- Если хост публичный, то к нему можно подключиться только с SSL
- Если хост непубличный, то к нему можно подключиться без SSL

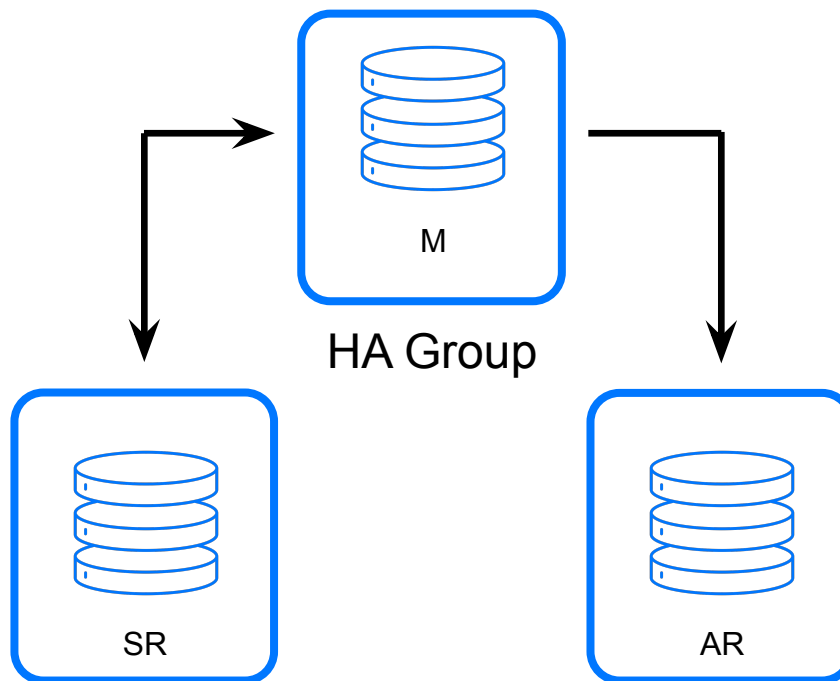


Как происходит восстановление после отказа?

Переключение на реплику 1

1. Реплика догоняет мастер.
2. Мастер переключается (до 30 секунд).
3. Обновляются клиенты:
 - примерно 3 секунды при клиентской балансировке,
 - до 30 секунд при использовании C-NAME.

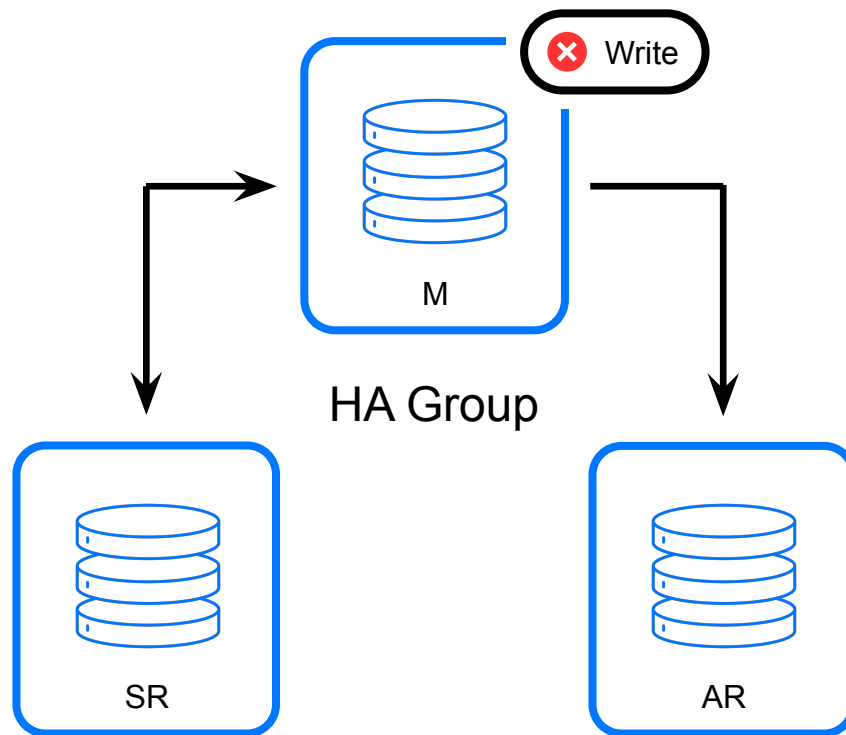
Итого: от 33 до 60 секунд
недоступности на запись.



Переключение на реплику 2

1. Реплика догоняет мастер.
2. Мастер переключается (до 30 секунд).
3. Обновляются клиенты:
 - примерно 3 секунды при клиентской балансировке,
 - до 30 секунд при использовании C-NAME.

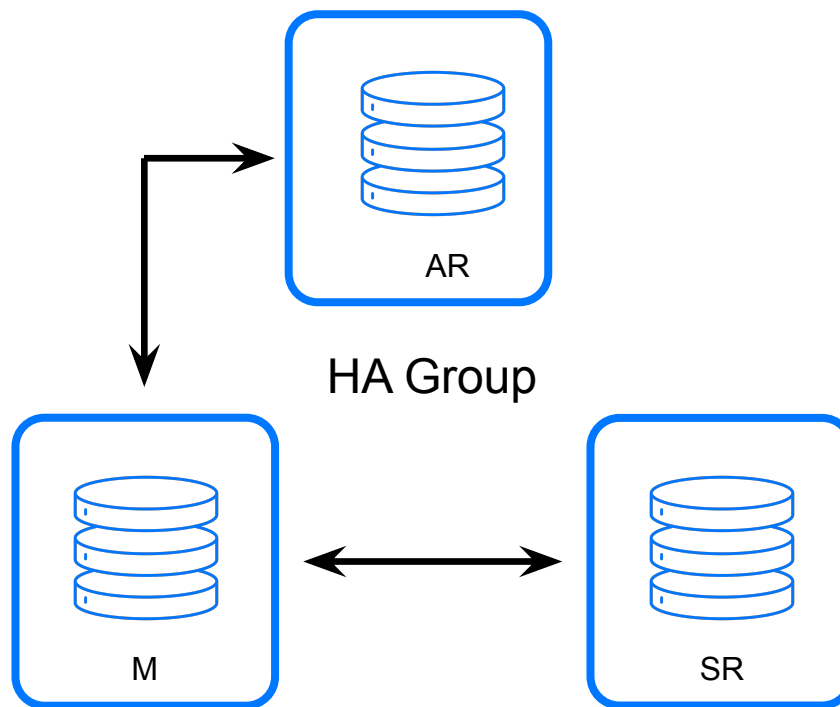
Итого: от 33 до 60 секунд
недоступности на запись.



Переключение на реплику 3

1. Реплика догоняет мастер.
2. Мастер переключается (до 30 секунд).
3. Обновляются клиенты:
 - примерно 3 секунды при клиентской балансировке
 - до 30 секунд при использовании C-NAME.

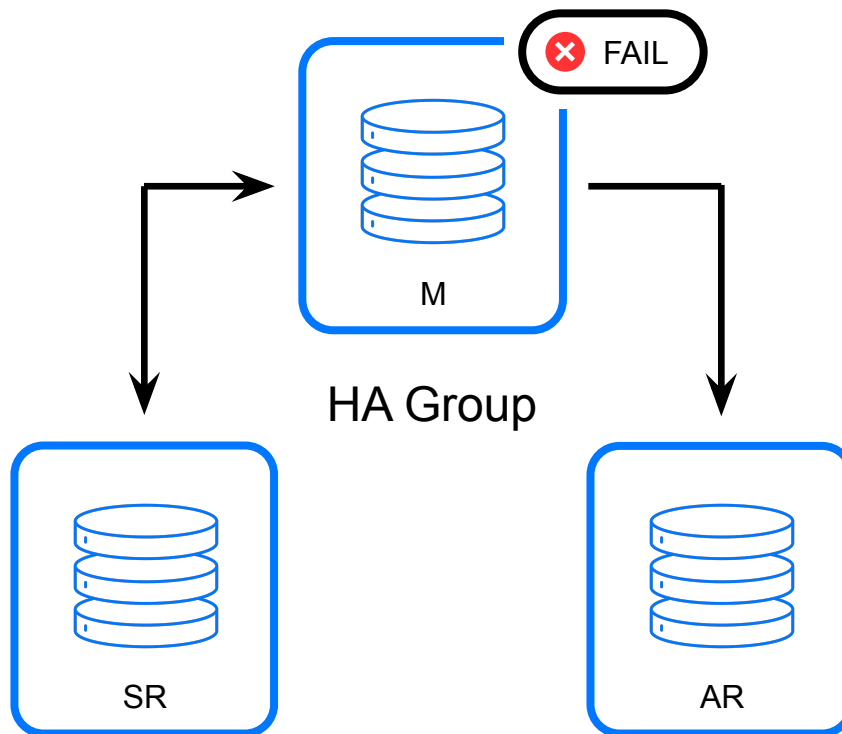
Итого: от 33 до 60 секунд
недоступности на запись.



Восстановление мастера 1

1. Происходит, когда мастер недоступен 30 секунд.
2. Восстановление делается на самую свежую реплику.
3. Обновляются клиенты.

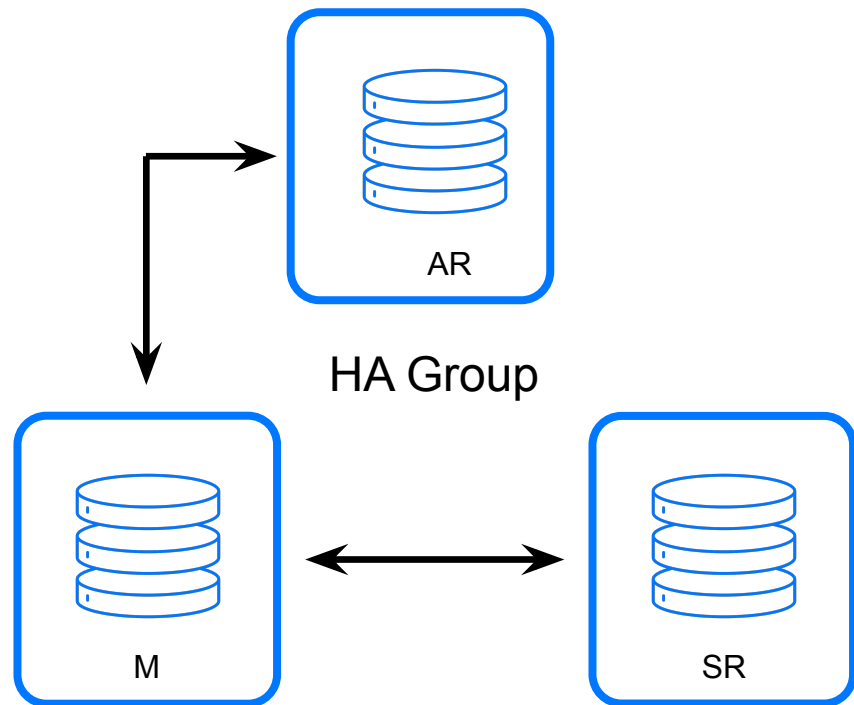
Итого: от 33 до 60 секунд
недоступности на запись.



Восстановление мастера 2

1. Происходит, когда мастер недоступен 30 секунд.
2. Восстановление делается на самую свежую реплику.
3. Обновляются клиенты.

Итого: от 33 до 60 секунд
недоступности на запись.





Экплуатация кластера

Окружения

Prestable

- Окружение, куда сначала добавляются все изменения в сервисе MDB
- Минорные обновления приходят туда первыми
- В первую очередь это касается внутренних утилит MDB
- Не рекомендуется для продакшн-сред

Production

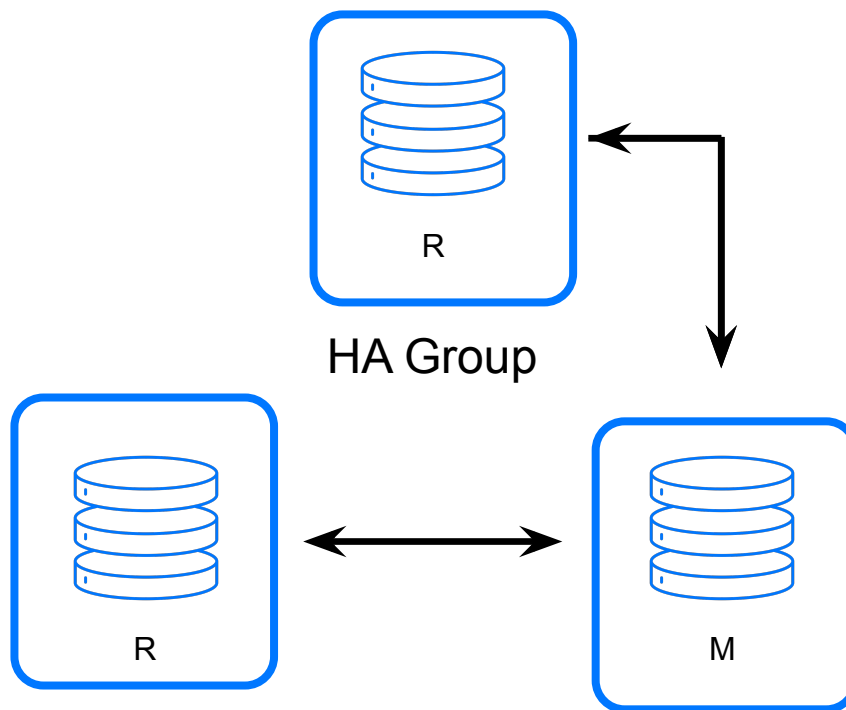
- В эту среду любые изменения приходят после того, как они какое-то время работали в окружении Prestable
- Рекомендуется для продакшн-сред

Изменение ресурсов сервиса Yandex Compute Cloud

Изменение vCPU, RAM,
размера диска:

1. Одна из реплик отключается.
2. Реплика изменяет размер.
3. Делается переключение на нее.
4. Остальные хосты меняются по аналогии с первым.

Итого: от 33 до 60 секунд
недоступности на запись.



Обновления

- › Минорные: рассылка письма и перезапуск базы
- › Мажорные:
 - PostgreSQL — pg_upgrade
 - MySQL — новый пакет
- › Недоступность записи во время обновления
 - PostgreSQL — минуты
 - MySQL — как переключение
- › Обновление можно делать только на одну версию выше



Изменить PostgreSQL-кластер

Базовые параметры

Имя кластера ?

concourse

Описание ?

Версия

11

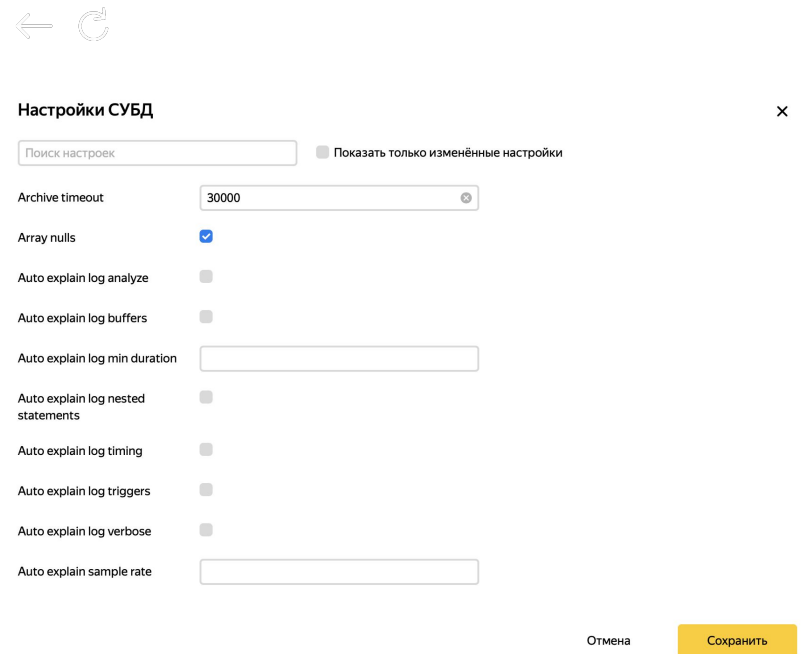
✓ 11

12

Класс хоста

Настройка кластера

- Кластер автоматически настраивает значения для базы по умолчанию
- Часто значения по умолчанию зависят от количества ядер или RAM
- Значения можно изменить (не выходя за рамки лимитов)
- Измененные значения останутся такими же при изменении количества vCPU кластера
- Некоторые изменения могут потребовать каскадного перезапуска баз в кластере (см. документацию СУБД)



Настройки СУБД

Поиск настроек ☐ Показать только изменённые настройки

Archive timeout	<input type="text" value="30000"/>
Array nulls	<input checked="" type="checkbox"/>
Auto explain log analyze	<input type="checkbox"/>
Auto explain log buffers	<input type="checkbox"/>
Auto explain log min duration	<input type="text"/>
Auto explain log nested statements	<input type="checkbox"/>
Auto explain log timing	<input type="checkbox"/>
Auto explain log triggers	<input type="checkbox"/>
Auto explain log verbose	<input type="checkbox"/>
Auto explain sample rate	<input type="text"/>

Отмена

Настройка max_connections

На всю СУБД зависит от:

- числа ядер (200 за 1 vCPU)
- типа CPU (частичное использование ядра или нет)
- типа базы (-15 соединений у управляемой PostgreSQL)

Для повышения значения надо увеличивать число ядер.

На пользователя:

- В сумме число соединений всех пользователей не должно превышать число на кластер



Настройки СУБД

Log statement	none	▼
Log temp files	-1	⊗
Maintenance work mem	67108864	⊗
Max connections	200	⊗
Max locks per transaction	64	⊗
Max parallel workers	8	⊗
Max parallel workers per gather	2	⊗
Max pred locks per transaction	64	⊗
Max prepared transactions	0	⊗
Max standby streaming delay	30000	⊗
Max wal size	4294967296	⊗

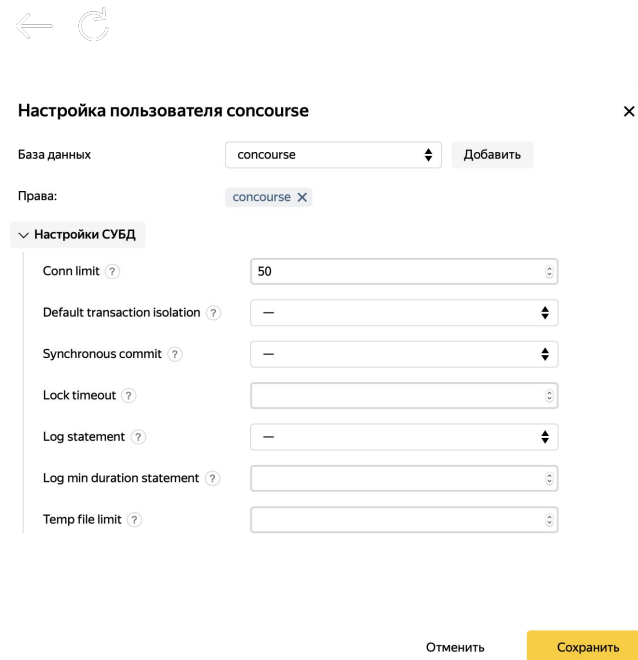


Отмена

Сохранить

Базы данных и пользователи

- Базы данных и пользователи создаются и удаляются только через API
- При создании базы надо указать пользователя-владельца
- Остальным пользователям надо выдать привилегии в базе, используя SQL



← ↻

Настройка пользователя concourse

База данных: Добавить

Права: concourse X

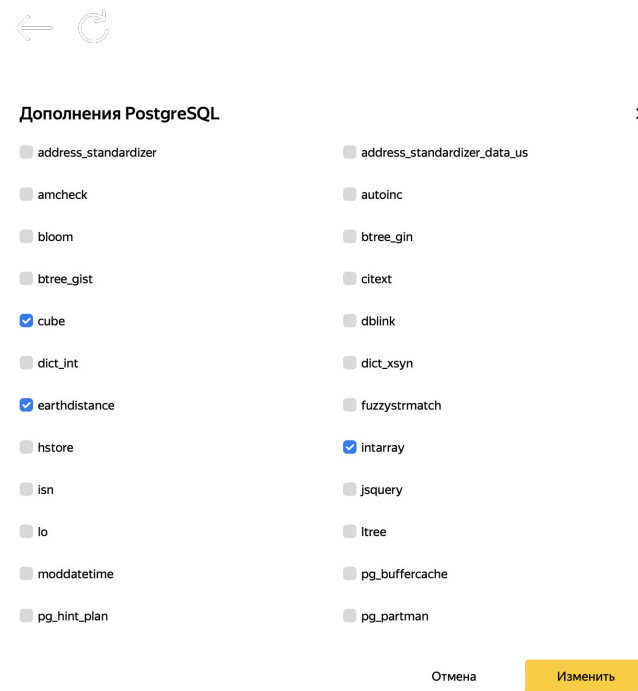
▼ Настройки СУБД

Conn limit ?	<input type="text" value="50"/>
Default transaction isolation ?	<input type="text" value="—"/>
Synchronous commit ?	<input type="text" value="—"/>
Lock timeout ?	<input type="text"/>
Log statement ?	<input type="text" value="—"/>
Log min duration statement ?	<input type="text"/>
Temp file limit ?	<input type="text"/>

Отменить Сохранить

Расширения PostgreSQL

- Выставляются через API
- Версии расширения меняются в зависимости от версии PostgreSQL



← ↺

Дополнения PostgreSQL

<input type="checkbox"/> address_standardizer	<input type="checkbox"/> address_standardizer_data_us
<input type="checkbox"/> amcheck	<input type="checkbox"/> autovacuum
<input type="checkbox"/> bloom	<input type="checkbox"/> btree_gin
<input type="checkbox"/> btree_gist	<input type="checkbox"/> citext
<input checked="" type="checkbox"/> cube	<input type="checkbox"/> dblink
<input type="checkbox"/> dict_int	<input type="checkbox"/> dict_xsyn
<input checked="" type="checkbox"/> earthdistance	<input type="checkbox"/> fuzzystmatch
<input type="checkbox"/> hstore	<input checked="" type="checkbox"/> intarray
<input type="checkbox"/> isn	<input type="checkbox"/> jsquery
<input type="checkbox"/> lo	<input type="checkbox"/> ltree
<input type="checkbox"/> moddatetime	<input type="checkbox"/> pg_buffercache
<input type="checkbox"/> pg_hint_plan	<input type="checkbox"/> pg_partman

Отмена Изменить

Health и Status

Health — состояние хоста / кластера — вывод системы мониторинга

- ALIVE
- DEGRADED
- DEAD
- UNKNOWN

Status — статус кластера, на который можно повлиять

- CREATING
- RUNNING
- ERROR — операция выполнена с ошибкой. Команда сервиса это увидит и будет исправлять.
В большинстве случаев это значит, что кластер работоспособен (нужно смотреть в Health), но через UI / CLI / API ничего с этим кластером сделать нельзя.
- UPDATING
- STOPPING
- STOPPED
- STARTING

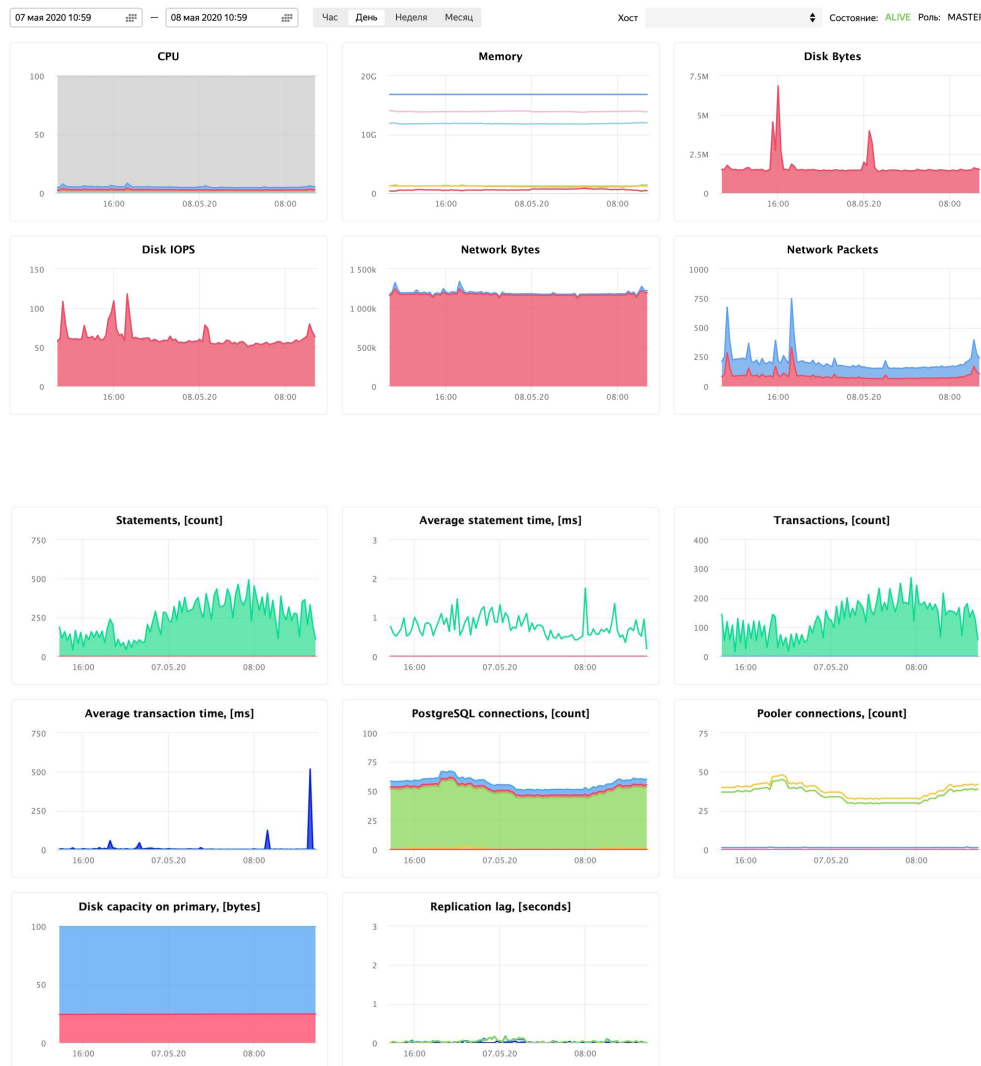
Мониторинг

Есть 2 разреза мониторинга:

- мониторинг всего кластера — там лежат метрики СУБД;
- мониторинг хостов — системные метрики хостов.

К метрикам можно добавить оповещения по e-mail или СМС с помощью сервиса Яндекс.Мониторинг.

В планах интеграция Яндекс.Мониторинга и Prometheus — она позволит добавить метрики баз в свой Prometheus.



Логи

Вы можете просматривать логи СУБД и pooler (для PostgreSQL). Параметры отображение логов доступны в настройках СУБД.




Дата	Сообщение
08.05.2020, 10:42:55.520	<div><div>automatic vacuum of table "carrier-aggregator-service.carrier_aggregator_service.item": index scans: 1 pages: 0 removed, 4862 remain, 0 skipped due to pins, 0 skipped frozen tuples: 245 removed, 505719 remain, 0 are dead but not yet removable, oldest xmin: 40007241 buffer usage: 1833 hits, 5337 misses, 126 dirtied avg read rate: 3.750 MB/s, avg write rate: 0.089 MB/s system usage: CPU: user: 0.08 s, system: 0.16 s, elapsed: 11.11 s</div><div>(nodb) (nouser)</div></div> <pre>{ "application_name": "", "command_tag": "", "connection_from": "", "context": "", "database_name": "", "detail": "", "error_severity": "LOG", "hint": "", "hostname": "rc1b-1e", "internal_query": "", "internal_query_pos": "0", "location": "", "message": "automatic vacuum of table \"carrier-aggregator-service.carrier_aggregator_service.item\": index scans: 1 pages: 0 removed, 4862 remain, 0 skipped due to pins, 0 skipped frozen tuples: 245 removed, 505719 remain, 0 are dead but not yet removable, oldest xmin: 40007241 buffer usage: 1833 hits, 5337 misses, 126 dirtied avg read rate: 3.750 MB/s, avg write rate: 0.089 MB/s system usage: CPU: user: 0.08 s, system: 0.16 s, elapsed: 11.11 s", "process_id": "268367", "query": "", "query_pos": "0", "session_id": "5eb50d65.4184f", "session_line_num": "5", "session_start_time": "2020-05-08 10:42:29", "sql_state_code": "00000", "transaction_id": "0", "user_name": "", "virtual_transaction_id": "35/13546247" }</pre>




Резервное копирование



Как делается резервное копирование?

Часть первая:

- Раз в неделю делается полная резервная копия; 
- Раз в день делается инкрементальная резервная копия;
- Полная резервная копия делается с реплики;
- Резервная копия шифруется;
- Время копирования можно выставить в настройках.

Дополнительные настройки

Начало резервного копирования (UTC) 

 22 : 00 

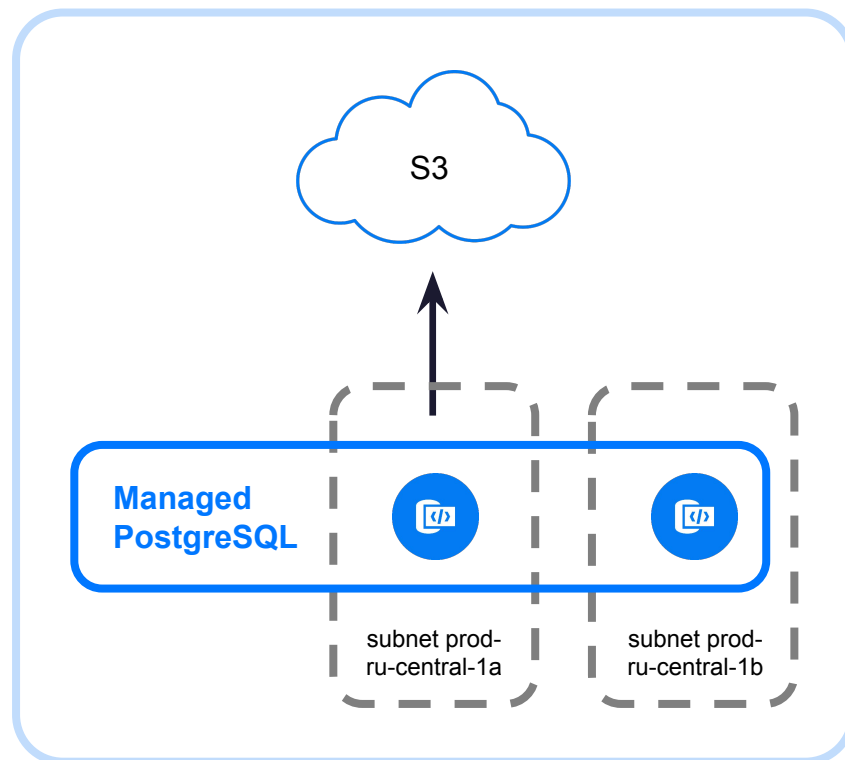
Доступ из DataLens 

Часть вторая:

- Транзакционные логи сжимаются и шифруются.

Хранение резервных копий

- Резервные копии хранятся в приватных бакетах S3
- К ним нет прямого доступа
- Сейчас длительность хранения — 7 дней
- Если требуется больше, то лучше делать дампы утилитами СУБД и хранить его у себя
- В планах — расширить длительность хранения с гибкими возможностями резервного копирования



Восстановление

- Доступно Point-in-Time Recovery за последние 7 дней
- Кластер восстанавливается целиком — со всеми базами, пользователями и всеми настройками
- При восстановлении создается новый кластер
- Самостоятельно сделанный дамп будет восстанавливаться без Point-in-Time Recovery



Восстановить PostgreSQL-кластер

Базовые параметры

Имя кластера ?	<input type="text" value="postgresql523"/>
Описание ?	<div>Restored from c9q6lbbrf1juaqu57h8m. Backup id: c9q6lbbrf1juaqu57h8m:base_000000010000 07E60000008C D 00000001000007D000000</div>
Окружение	<input type="text" value="PRODUCTION"/>
Версия	<input type="text" value="11"/>
Дата и время восстановления (UTC) ?	<input type="text" value="06.05.20 22:03"/>
Каталог	<input type="text" value="nrk-demo"/>



Доступность и SLA

Доступность приложения — совместная работа

Команда MDB

- Механизмы, обеспечивающие обнаружение и исправление сбоев
- Явно разделенные SLA на чтение и запись
- Мониторинг свободного места и закрытие записи, если нет свободного места (95% занято)

Пользователь

- Настройка клиентов для работы в HA-конфигурации
- Настройка приложений, позволяющая разделить чтение и запись
- Своевременное расширение свободного места

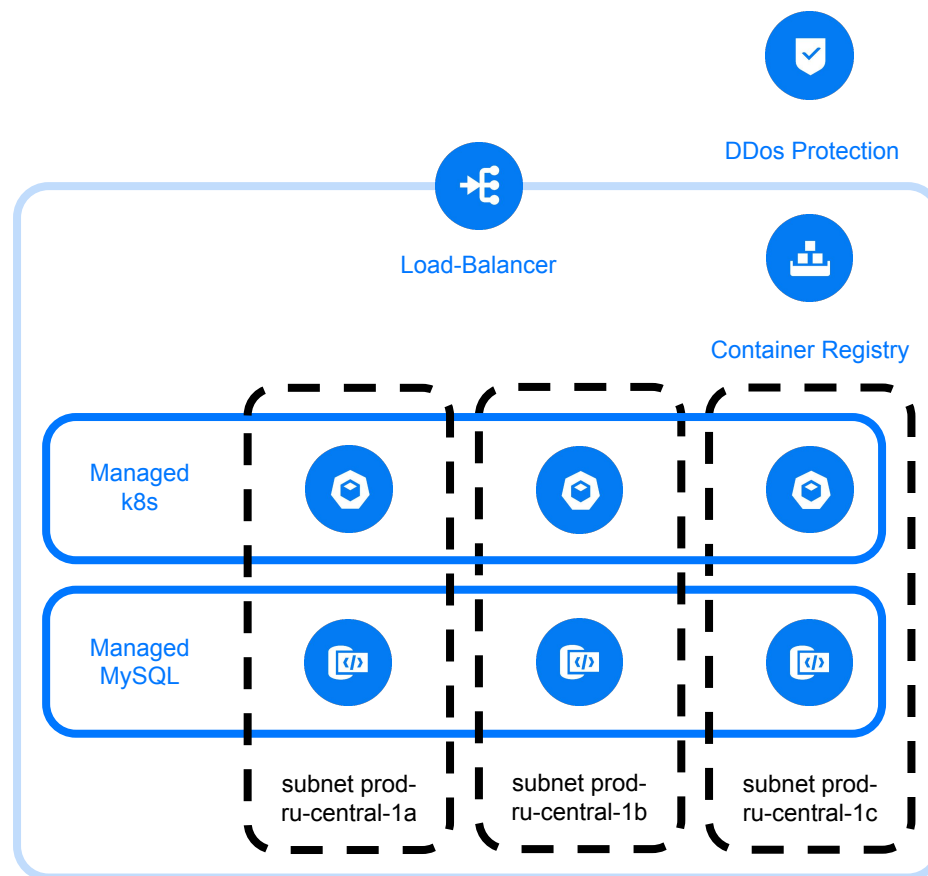
SLA

Uptime

- 99,99% на чтение
- 99,95% на запись

Распространяется на конфигурацию

- С минимум двумя хостами в разных зонах доступности
- При использовании поддерживаемых клиентов





КВОТЫ И ЛИМИТЫ



Поддерживаемые версии, осень 2021

PostgreSQL

- › 10
- › 10 для 1С
- › 11
- › 12
- › 13

MySQL

- › 5.7
- › 8.0

Поддерживаемые размеры, осень 2021

Минимальный размер

- 2 ядра 5% vCPU
- 2 GB RAM

Максимальный размер

- 80 ядра 100% vCPU
- 640 GB RAM

Типы VM

- Burstable — небольшие кластера для тестовых сред. Не рекомендуем для продакшн-среды.
- Standard — отношение vCPU к RAM: 1/4
- Memory-Optimized — отношение vCPU к RAM: 1/8

Новые размеры будут появляться в Yandex Compute Cloud

Диски, осень 2021

Сетевые диски с репликацией

- Доступны Network SSD, Network HDD, Network SSD non-replicated
- Наследуются все ограничения сервиса Yandex Compute Cloud*
- Можно изменять размер дисков
- Максимальный размер — 8 TB для non-replicated дисков, 4 TB для остальных

Нельзя поменять тип диска

- Доступны для любых конфигураций

cloud.yandex.ru/docs/compute/concepts/limits#limits-disks

Локальные SSD

- Локальные SSD-диски без ограничений производительности
- Доступны для конфигураций с 3 хостами в разных зонах доступности
- Максимальный размер — 1,5 TB
- Недоступно изменение размера дисков


Квоты сервиса

Квота общая на все управляемые базы данных

- PostgreSQL
- MySQL
- ClickHouse
- Redis
- MongoDB

Доступны в разделе квоты.

Увеличиваются по запросу в техническую поддержку.



Как переехать в управляемую базу

Этап 1. Подготовка базы

Необходимо заранее создать:

- Кластер
- Пользователей
- Базы
- Расширения

Далее сделать восстановление базы с помощью:

- pg_restore
- mysql



Этап 2. Тестирование приложений

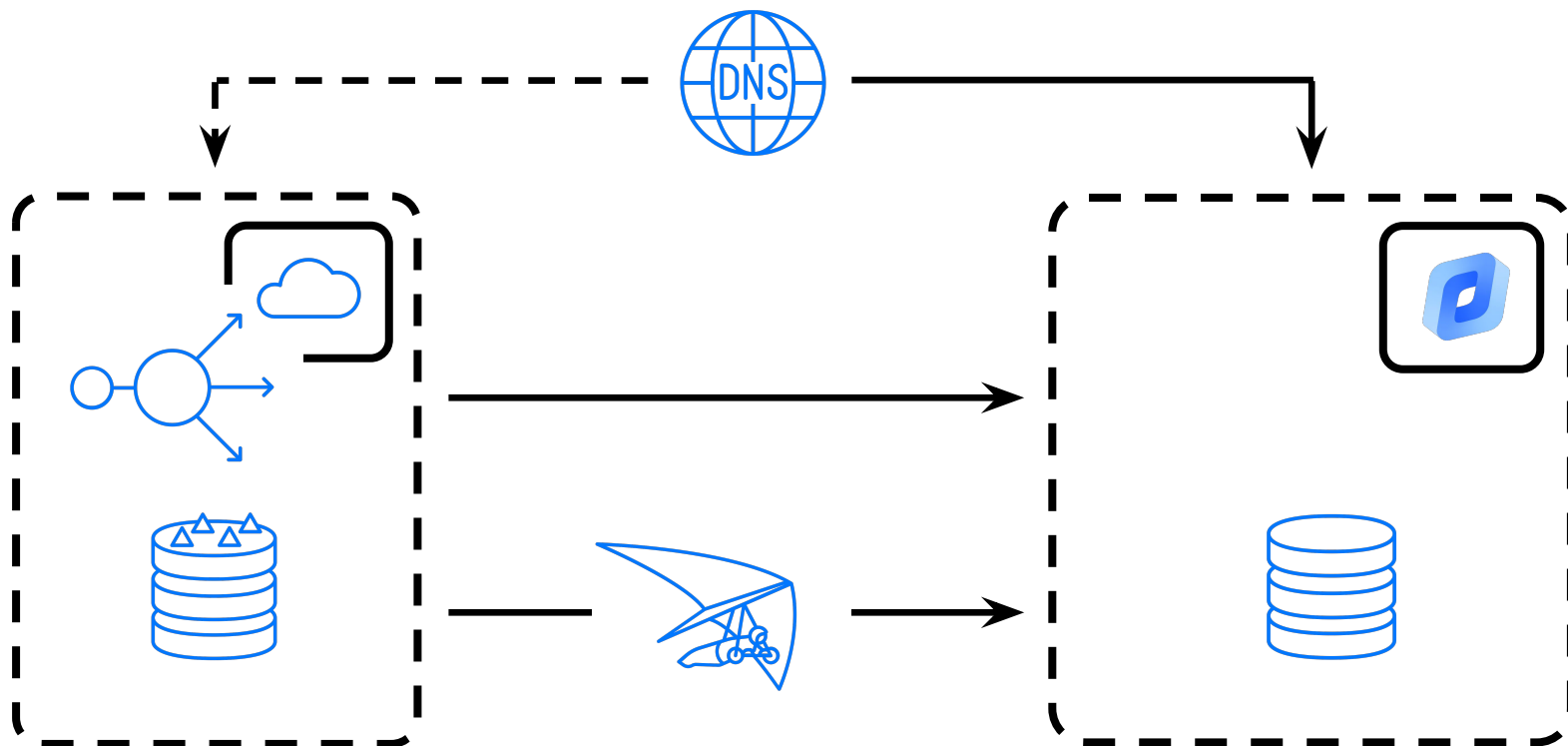
Разверните тестовый контур приложения

Проведите тесты

- Функциональные
- Нагрузочные

Убедитесь, что приложение и база работают корректно.

Этап 3. Миграция PROD. Простой вариант





Этап 3. Миграция PROD. Минимизируем простой

- › Сервис Yandex DataTransfer для бесшовной миграции с MySQL и PostgreSQL



Итоги

Итоги

Сегодня мы:

- поняли как работать с управляемыми СУБД.



Дополнительные материалы по работе с Yandex.Cloud

- [Yandex Managed Service for PostgreSQL](#)
- [Yandex Managed Service for MySQL®](#)
- [Data Transfer](#)



Домашнее задание

Домашнее задание

Давайте посмотрим ваше [домашнее задание](#).

- Вопросы по домашней работе задавайте **в чате** мессенджера Slack.
- Задачи можно сдавать **по частям**.
- Зачёт по домашней работе проставляется после того, как **приняты все задачи**.

**Задавайте вопросы и
пишите отзыв о лекции!**

Нарек Татевосян