



R.Style Softlab

TARANTOOL

Tarantool Cartridge: разработка отказоустойчивого кластера Григоров Андрей (R-Style Softlab) репекsglazami

Пилотный проект

Система аутентификации интернет-банка



TARANTOOL

План доклада

- Что такое Tarantool?
- Варианты хранения данных: в памяти и на диске
- Шардирование данных в Tarantool
- Как построить кластер Tarantool
- Фреймворк Tarantool Cartridge
- > Разработка приложений на java, работающих с Tarantool Cartridge
- **Конкурс**

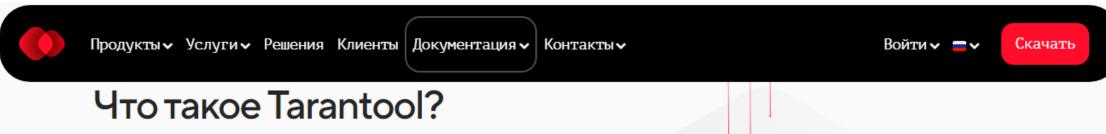
Онлайн-опрос

Заходим на menti.com

39 20 89 9

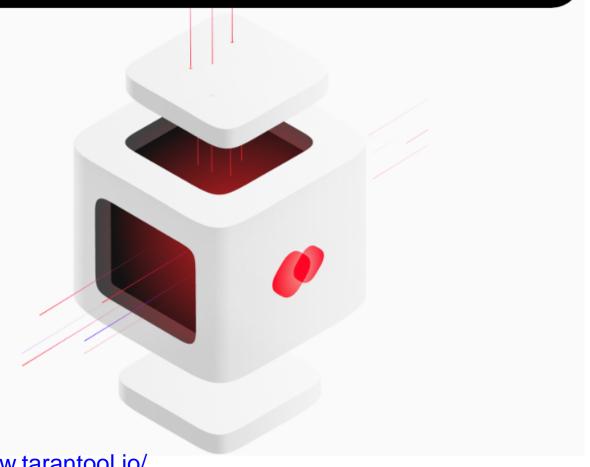


Что такое Tarantool?



Tarantool — это база данных с открытым исходным кодом, которая умеет хранить все в оперативной памяти. Она обслуживает до миллиона запросов в секунду, ищет по вторичным индексам и поддерживает SQL.

В Tarantool можно исполнять код рядом с данными. Встроенный язык программирования Lua поможет реализовать любую бизнес-логику и ускорить ваше решение. Избавляйтесь от устаревших записей, синхронизируйтесь с другими источниками данных, реализуйте HTTP-сервис.



https://www.tarantool.io/

История создания Tarantool

2008 год - Tarantool появился в рамках проекта социальной сети «Мой мир» как замена key-value хранилищу memcached для хранения сессий и профилей пользователей

📵 мой мир

В дальнейшем стал использоваться на проектах Mail.Ru и за её пределами (Badoo, Meraфoн, Аэрофлот, Альфа-банк) для обработки «горячих» данных: пользовательские онлайн-сессий, настройки онлайн-приложений, различные кеши и т.д.

2016 год - Исходный код Tarantool опубликован в открытом доступе под лицензией BSD.

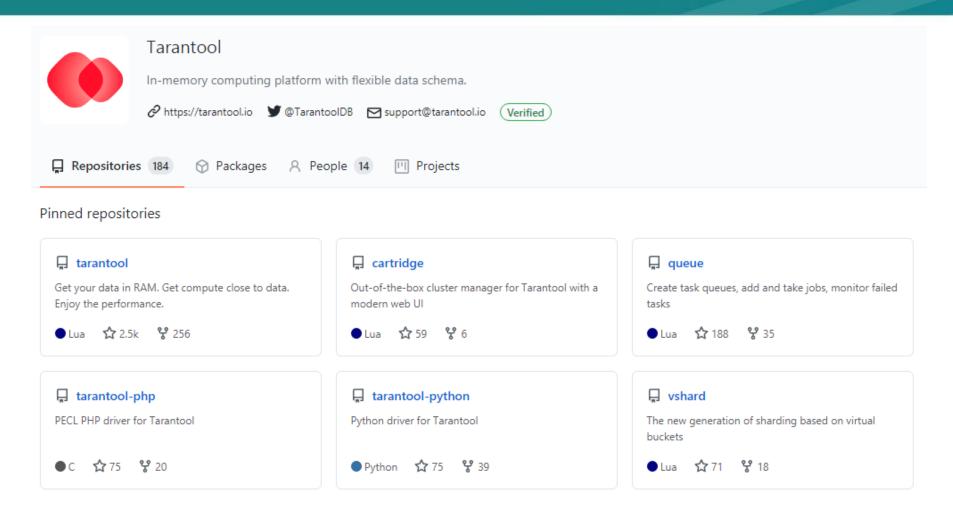
Версия 1.7 (2016 год) - Добавлен движок хранения данных на флеш и жестких дисках - Vinyl

Версия 2.1 (2018 год) - Добавлена поддержка SQL

Версия 2.1.2 (2019 год) - Фреймворк для разработки кластерных приложений (Tarantool Cartridge) стал частью Tarantool с открытым кодом

Версия 2.6 (2020 год) - Синхронная репликация данных

Tarantool на GitHub



https://github.com/tarantool

Отличия от других NoSQL-решений



https://www.youtube.com/watch?v=sbSTZmkk5LE

- Модель данных
- Кооперативный транзакционный менеджер
- > Сервер приложений

Сервер приложений Lua

print('Hello, world!')

```
app.lua
```

```
$ tarantool app.lua
Hello, world!
```



Lua (лу́а, с порт. — «луна») — скриптовый язык программирования, разработанный в подразделении Tecgraf (Computer Graphics Technology Group) Католического университета Рио-де-Жанейро (Бразилия). Интерпретатор языка является свободно распространяемым, с открытыми исходными текстами на языке Си.

https://ru.wikipedia.org/wiki/Lua

Сервер приложений Lua

base.lua

```
box.cfq {
   listen = 3301
box.once("bootstrap",
       function()
            box.schema.space.create('users')
            box.space.users:format({
                { 'login', 'string' },
                { 'password', 'string' }
            })
            box.space.users:create index('user login index', {
                parts = { field = "login", type = "string" } },
                if not exists = true
            })
        end
```

```
$ tarantool base.lua
2021-02-25 15:00:41.250 [41436] main/103/base.lua C> version 2.5.3-0-gf93e480
2021-02-25 16:00:44.250 [41436] main/103/base.lua C> log level 5
...
```

Модель данных. Отличия от других

Отличие от реляционных СУБД Oracle, MySQL, PostgreSQL	Можно не иметь схему данных
Отличие от документоориентированных СУБД	Может быть схема данных
Couchbase, MongoDB	Схема данных не хранится в
	каждом документе

Модель данных

Tarantool - документоориентированная СУБД

Модель данных. Возможности Tarantool

- Управление таблицами/спейсами
 Транзакции, вторичные ключи, локальные индексы, функциональные индексы
- Обеспечение корректности данных Возможность задания схемы
- Эволюция схемы данных
 Онлайн DDL (добавление/удаление индексов неблокирующая операция)
 Можно налету включать/выключать проверку форматов данных
- Управление размером данных Компрессия данных

Модель данных. Пример описания модели данных

Модель данных. Пример описания модели данных

```
local users = box.schema.space.create('users',
    { if not exists = true }
users:format({
                                      users:create index('user login index', {
    { 'bucket id', 'unsigned' },
                                          parts = { field = "login", type = "string" } },
    { 'uuid', 'string' },
                                           if not exists = true
    { 'login', 'string' },
                                       })
   { 'password', 'string' },
   { 'status', 'string' }
                                       users:create index('user uuid index', {
})
                                          parts = { field = "uuid", type = "string" } },
                                           if not exists = true
                                       })
```

Модель данных. Пример описания модели данных

```
local users = box.schema.space.create('users',
    { if not exists = true }
users:format({
                                      users:create index('user login index', {
    { 'bucket id', 'unsigned' },
                                          parts = { field = "login", type = "string" } },
    { 'uuid', 'string' },
                                           if not exists = true
    { 'login', 'string' },
                                       })
   { 'password', 'string' },
   { 'status', 'string' }
                                       users:create index('user uuid index', {
})
                                          parts = { field = "uuid", type = "string" } },
                                           if not exists = true
                                       })
```

```
box.space.users:insert { 1, '123e4567-e89b-12d3-a456-426655440000', 'misterX',
'c4ca4238a0b923820dcc509a6f75849b', 'ACTIVE' }
```

Семейства подсистем хранения данных

Постраничные

Page-oriented storage engine

Журналированные

Log-structured storage engine

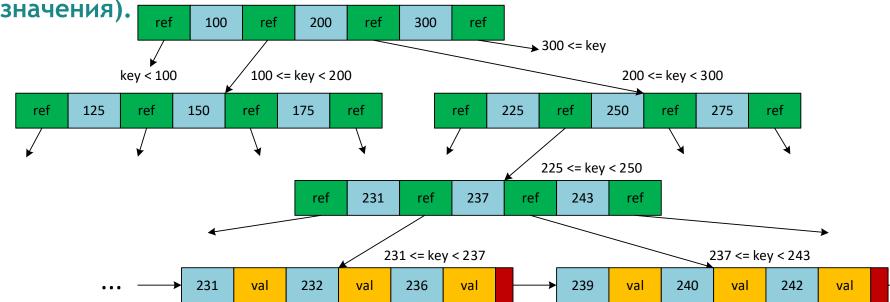
Постраничные системы хранения данных

- Использует индексная структура на основе В-дерева.
- Использование В-деревьев впервые было предложено Р. Бэйером (R. Bayer) и Э. МакКрейтом (E. McCreight) в 1970 году в лаборатории Boeing.

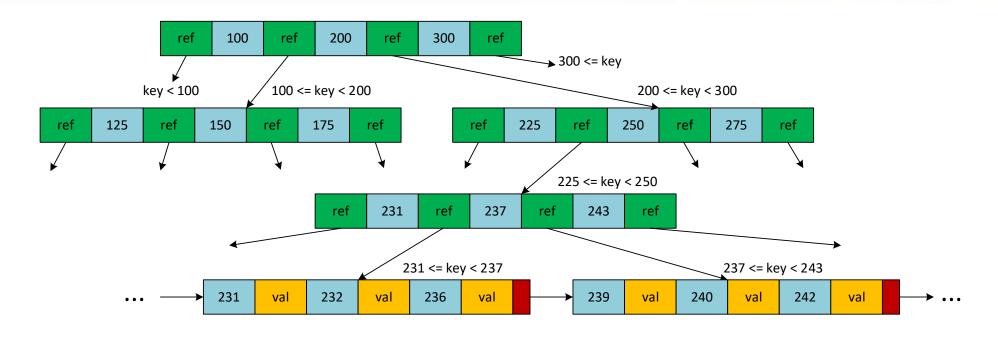
B+tree

- В-дерево разбивают БД на страницы фиксированного размера (обычно 4-16 Кб).
- Страница читается и записывается полностью.
- Страницы имеют адрес, по которым на них могут ссылаться другие страницы.
- Одна страница назначается корнем В-дерева. С неё начинается поиск.
- Каждая не листовая страница содержит список ключей, которые определяют границы диапазонов ключей, и ссылки на дочерние страницы, которые соответствуют этим диапазонам.

Страницы-листья содержат отдельные ключи и их значения (или ссылки на страницы, которые содержат значения).

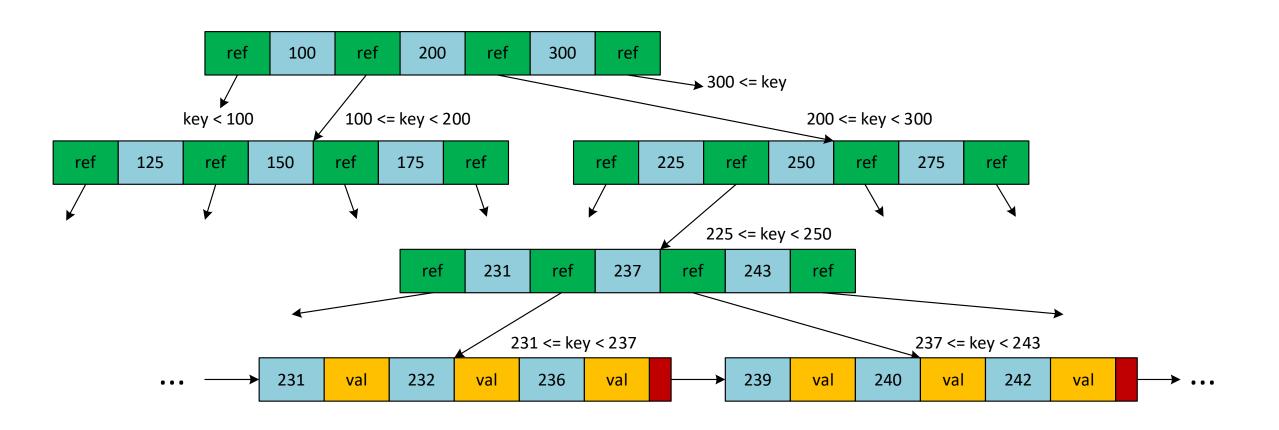


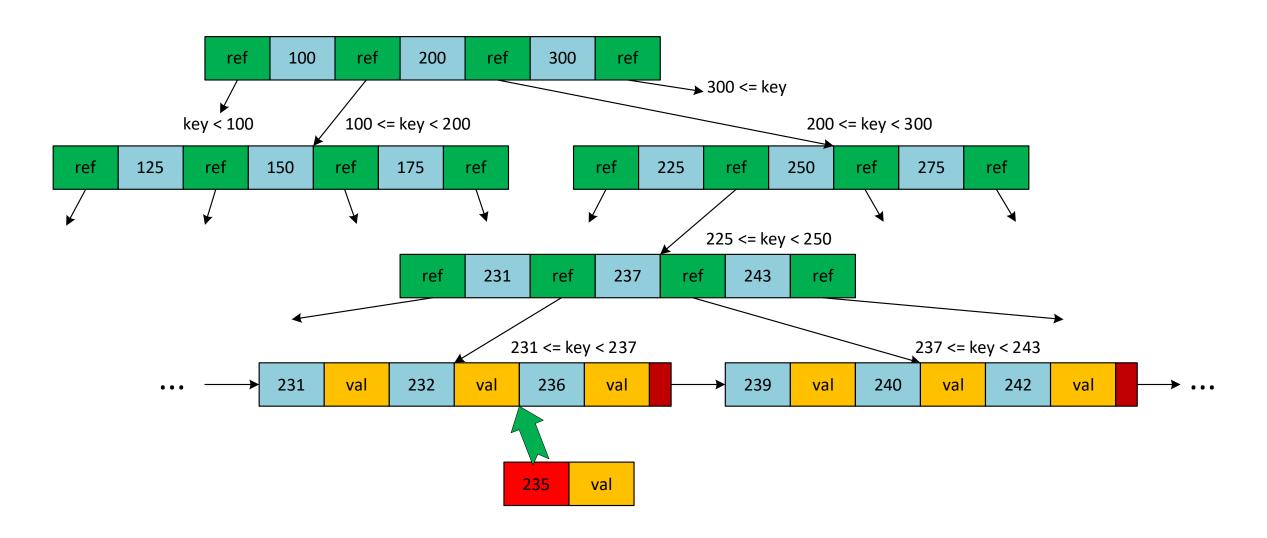
Свойства B+tree

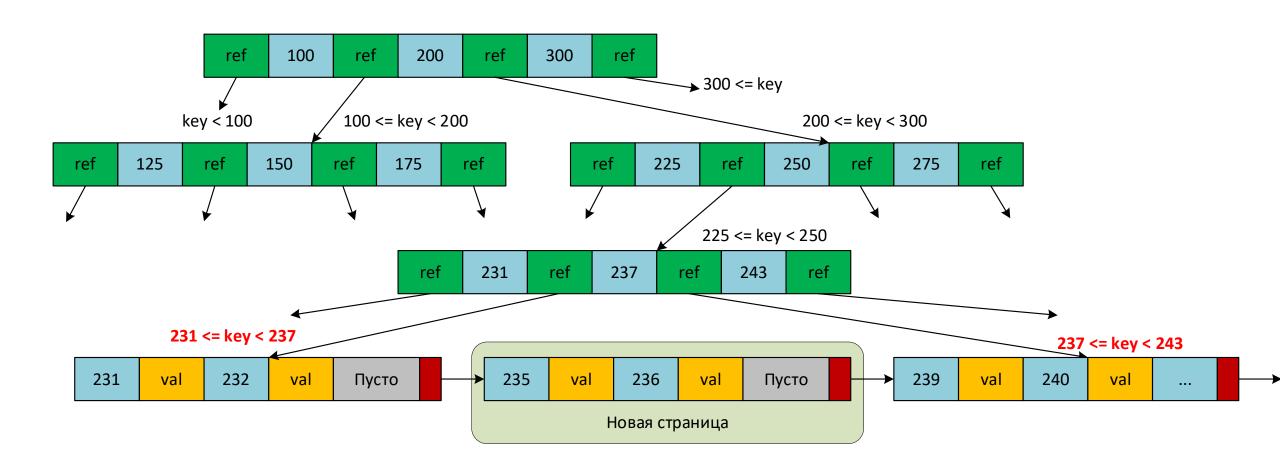


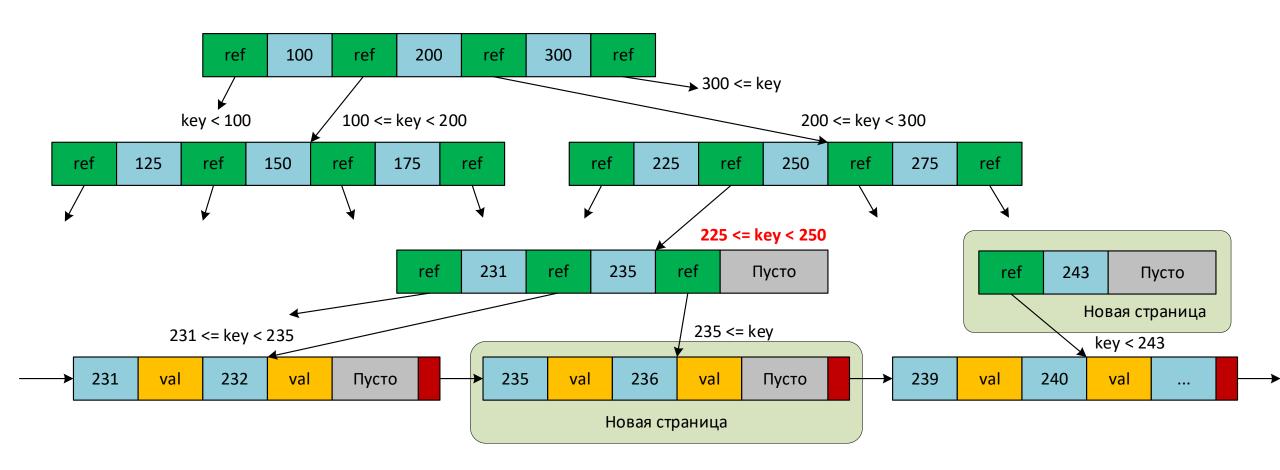
- Свойства В-дерева:
 - ➤ сбалансированное
 - **>** сильноветвистое

Операция	Среднее время	Худшее время
Поиск элемента	O(logN)	O(logN)
Вставка элемента	O(logN)	O(logN)
Удаление элемента	O(logN)	O(logN)











Your PC ran into a problem and needs to restart. We're just collecting some error info, and then we'll restart for you.

20% complete



For more information about this issue and possible fixes, visit https://www.windows.com/stopcode

If you call a support person, give them this info:

Stop code: CRITICAL PROCESS_DIED

В-дерево. Обеспечение надёжности

Журнал упреждающей записи (write-ahead log) – файл, в который последовательно записываются все модификации В-дерева до того, как эти модификации в реальности будут примерены к страницам дерева.

WAL B Tarantool

Режим работы журнала упреждающей записи в Tarantool определяется значением настройки wal_mode.

Возможные значения:

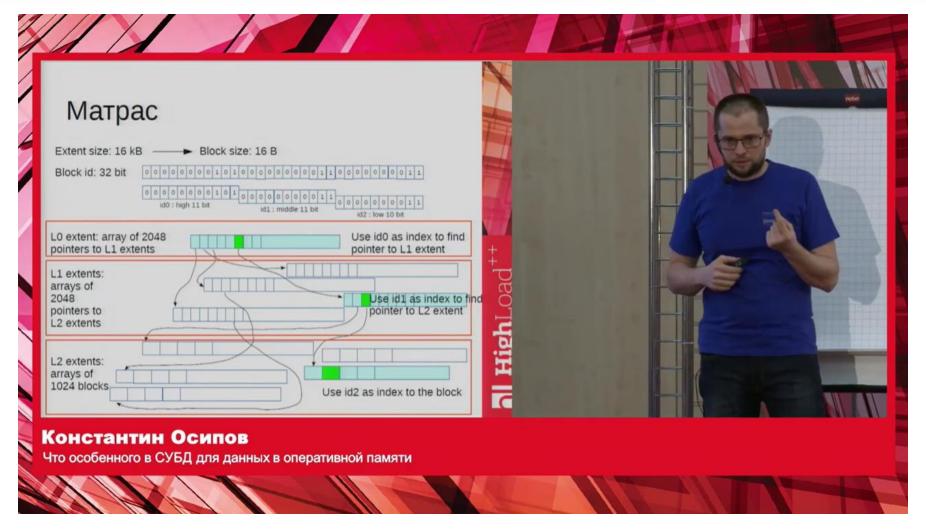
- none журнал упреждающей записи не используется;
- write файберы ожидают записи в журнал упреждающей записи (не fsync(2)); используется по умолчанию;
- fsync файберы ожидают данные, синхронизация fsync(2) следует за каждой операцией write(2).

Memtx. Используются B+*tree

Свойства bps-tree в memtx:

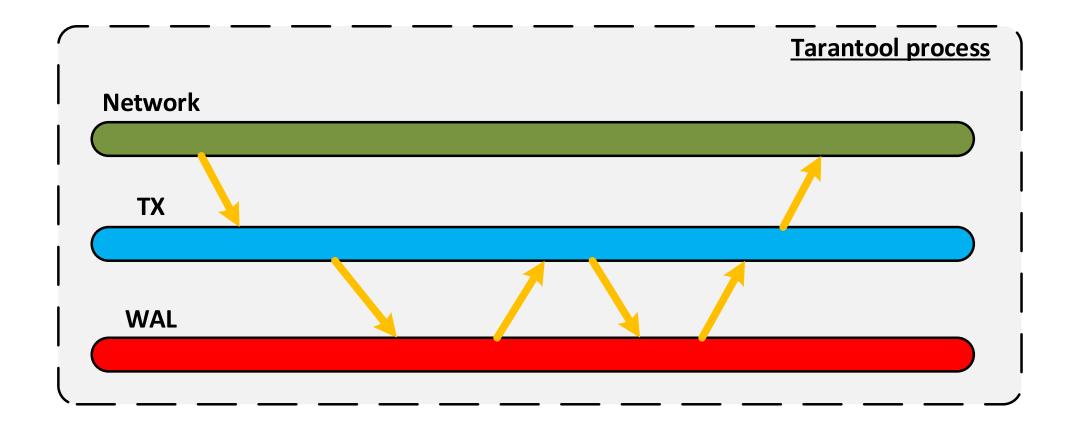
- Компактное хранение данных и преимущества cache-oblivious (от Вtree)
- > Дешёвая итерация по списку ключей (от B+-tree)
- Большая компактность и сбалансированность (от B*-tree)
- У Использование matras для MVCC и аллокации памяти

Memtx. Для чего нужен матрас?



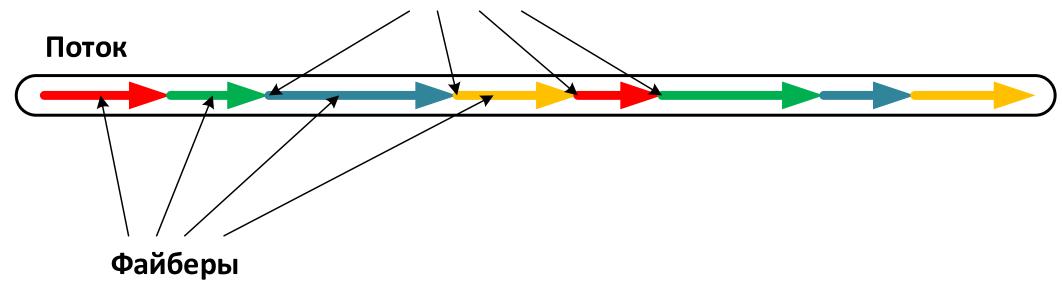
https://www.youtube.com/watch?v=yrTF3qH8ey8

Кооперативная многозадачность



Кооперативная многозадачность

Точки передачи управления



Плюсы:

- Поток всегда занят исполнением задач (эффективная утилизация CPU)
- Отсутствие проблем с кешем CPU

Минусы:

Нет возможности использовать все ядра процессора

Транзакции в memtx в версиях 2.5 и ниже

```
function transferMoney(from_account_id, to_account_id, amount_of_money)
    box.begin()
    box.space.accounts:update(from_account_id, { { '-', BALANCE_FIELD_NUM, amount_of_money } })
    box.space.accounts:update(to_account_id, { { '+', BALANCE_FIELD_NUM, amount_of_money } })
    box.commit()
end
```

Что мы знаем о транзакциях в memtx Tarantool версии 2.5 и ниже?

- После начала транзакции box.begin() все операции записи не приводят к записи в WAL, они копятся до вызова box.commit. Это позволяет не передавать управление другим файберам.
- Транзакции атомарны.
- Уровень изоляции транзакций serializable, если нет проблем с записью в WAL.
- Если происходит сбой при записи в WAL, то выполняется откат транзакции при этом другие транзакции могут прочитать неподтверждённые изменения (dirty read), то есть уровень изоляции транзакций становится read uncommitted.
- > Выполнение yield-операции в файбере приводит к откату транзакции.
- Не поддерживаются интерактивные транзакции.

Новый менеджер транзакций в v.2.6.1

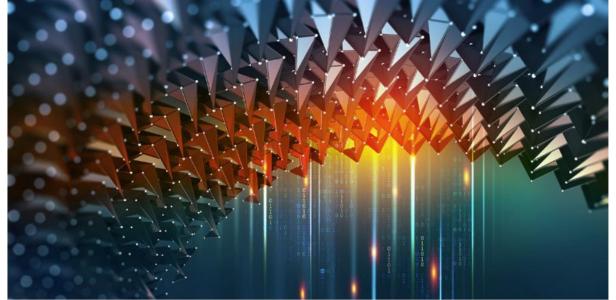


https://www.youtube.com/watch?v=BiF7L2id-TU

🗼 alexlyapunov 8 февраля 2021 в 16:45

Менеджер транзакций для базы данных в оперативной памяти

Блог компании Mail.ru Group, Высокая производительность, Алгоритмы, Хранение данных, Tarantool



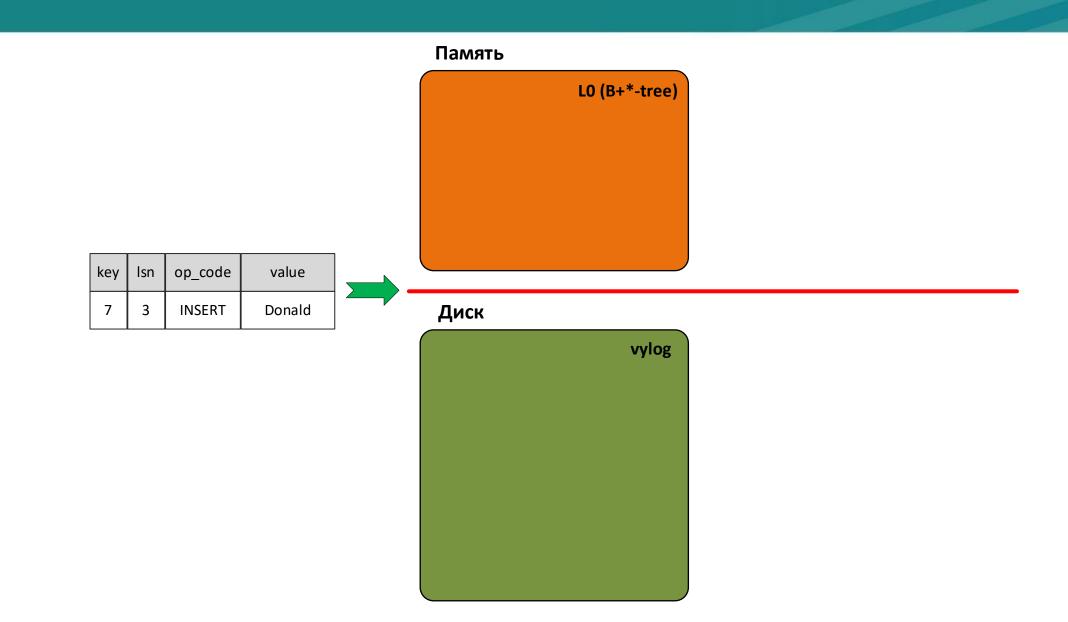
https://habr.com/ru/company/mailru/blog/540842/

Vinyl. Движок для хранения данных на диске

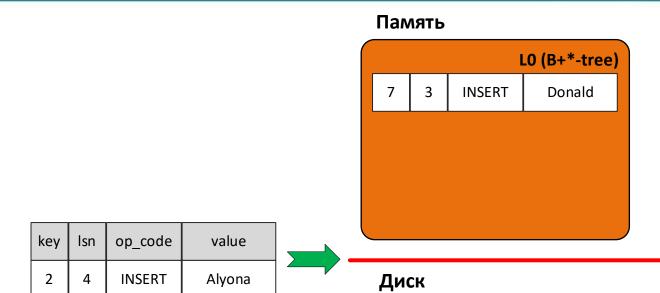
- Появился в версии 1.7.
- > В качестве основной структуры данных для хранения используются LSM-деревья (Log Structured Merge Tree).



Vinyl. LSM-дерево



Vinyl. LSM-дерево





Память

		ı	_0 (B+*-tree)
2	4	INSERT	Alyona
7	3	INSERT	Donald

key	Isn	op_code	value
1	6	INSERT	Anton

			vylog
7	3	INSERT	Donald
2	4	INSERT	Alyona

Память

			L0 (B+*-tree)
1	6	INSERT	Anton
2	4	INSERT	Alyona
7	3	INSERT	Donald
7	3	INSERT	Donald

key	Isn	op_code	value
7	10	REPLACE	Joe

			vylog
7	3	INSERT	Donald
2	4	INSERT	Alyona
1	6	INSERT	Anton

Память

		ı	L0 (B+*-tree)
1	6	INSERT	Anton
2	4	INSERT	Alyona
7	10	REPLACE	Joe

key	Isn	op_code	value
7	17	DELETE	

			vylog
7	3	INSERT	Donald
2	4	INSERT	Alyona
1	6	INSERT	Anton
7	10	REPLACE	Joe

Память

			LO (B+*-tree)
1	6	INSERT	Anton
2	4	INSERT	Alyona
7	17	DELETE	

key	Isn	op_code	value
8	18	INSERT	Stepan

			vylog
7	3	INSERT	Donald
2	4	INSERT	Alyona
1	6	INSERT	Anton
7	10	REPLACE	Joe
7	17	DELETE	

Память

		l	LO (B+*-tree)
1	6	INSERT	Anton
2	4	INSERT	Alyona
7	17	DELETE	
8	18	INSERT	Stepan

			vylog
7	3	INSERT	Donald
2	4	INSERT	Alyona
1	6	INSERT	Anton
7	10	REPLACE	Joe
7	17	DELETE	
8	18	INSERT	Stepan

Память

L0 (B+*-tree)			
1	6	INSERT	Anton
2	4	INSERT	Alyona
7	17	DELETE	
8	18	INSERT	Stepan

			vylog
7	3	INSERT	Donald
2	4	INSERT	Alyona
1	6	INSERT	Anton
7	10	REPLACE	Joe
7	17	DELETE	
8	18	INSERT	Stepan

			run
1	6	INSERT	Anton
2	4	INSERT	Alyona
7	17	DELETE	
8	18	INSERT	Stepan

Память





			run
1	6	INSERT	Anton
2	4	INSERT	Alyona
7	17	DELETE	
8	18	INSERT	Stepan

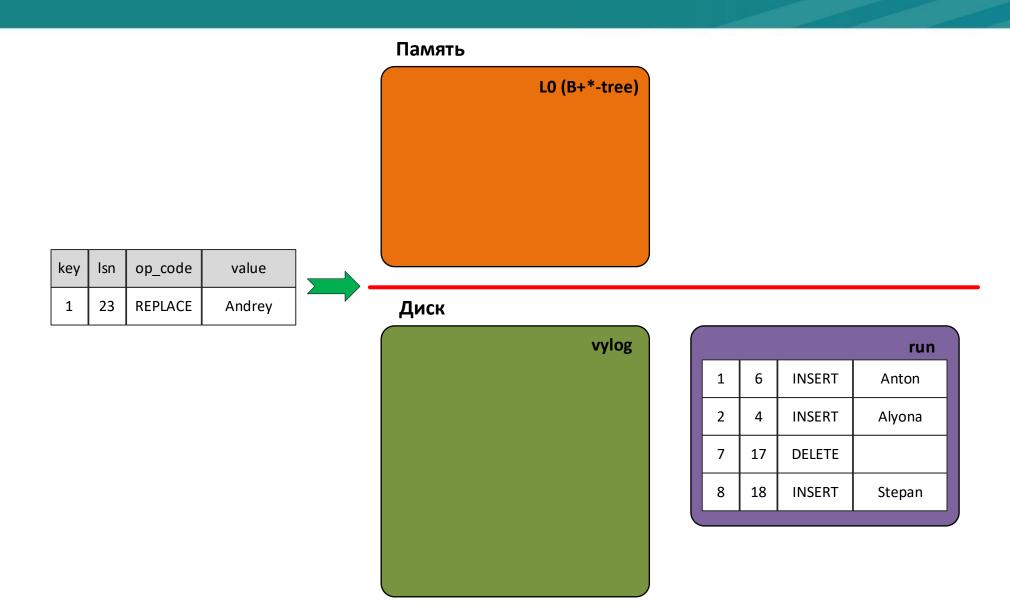
Память

LO (B+*-tree)

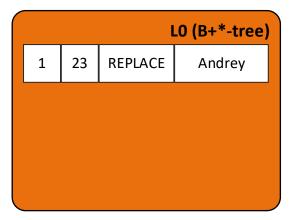
Диск

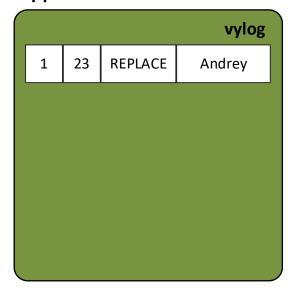
vylog

			run
1	6	INSERT	Anton
2	4	INSERT	Alyona
7	17	DELETE	
8	18	INSERT	Stepan

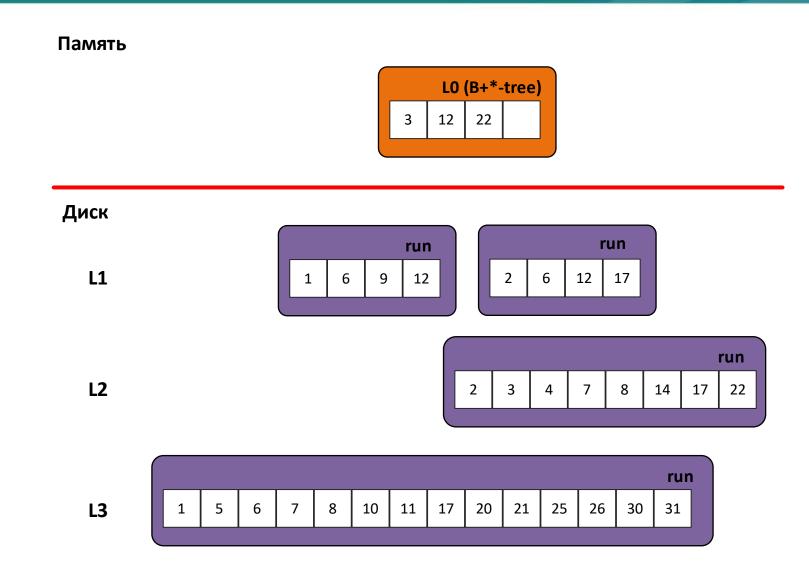


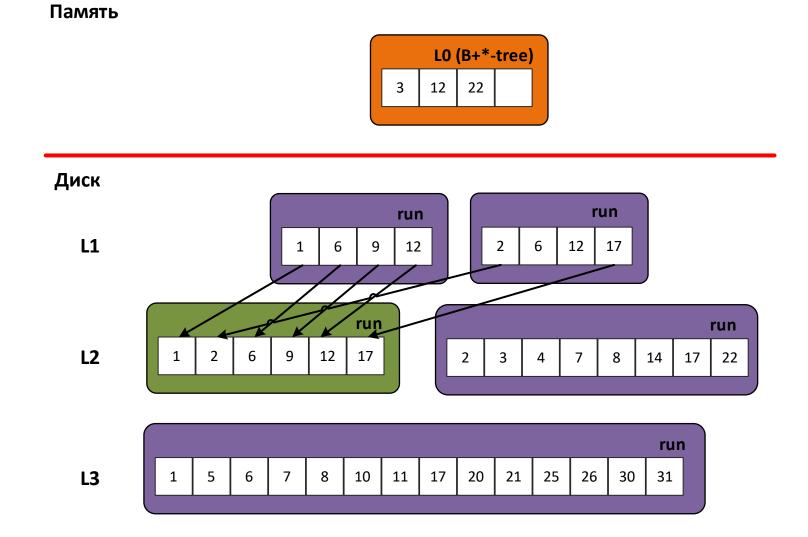
Память

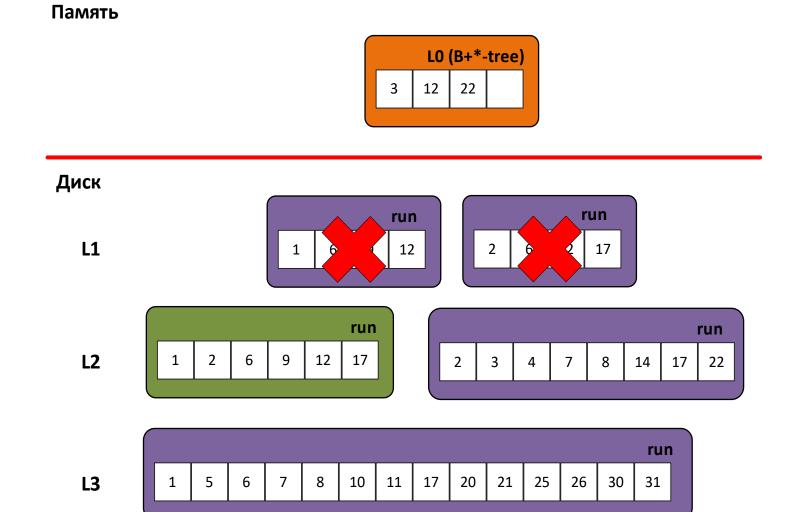




			run
1	6	INSERT	Anton
2	4	INSERT	Alyona
7	17	DELETE	
8	18	INSERT	Stepan







Память



Диск

L1

1 2 6 9 12 17



26

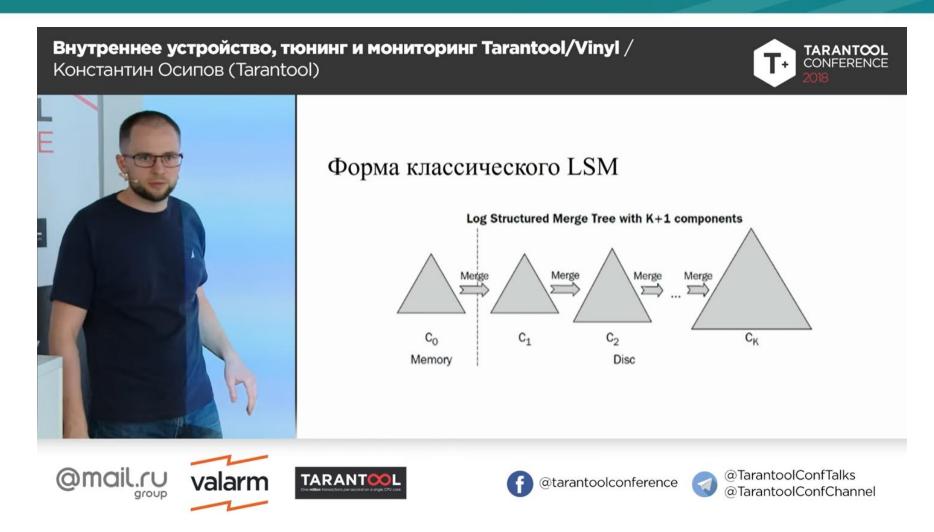
30

run

31

L3 1 5 6 7 8 10 11 17 20 21 25

Vinyl. Внутреннее устройство, тюнинг и мониторинг



https://www.youtube.com/watch?v=QrNTmBdQMrU

Свойства LSM-деревьев

Преимущества

- Обычно LSM-деревья, обеспечивают большую пропускную способность на запись, чем В-деревья.
- > Усиление записи (write amplification) обычно ниже, чем в В-деревьях.
- LSM-деревья сжимаются лучше, чем В-деревья. Файлы с данными занимают меньше места на диске.

Недостатки

- Один ключ может дублироваться в разных файлах с SSTable.
- Непредсказуемая скорость чтения.
- Скорость ответа зависит от фоновых задач.

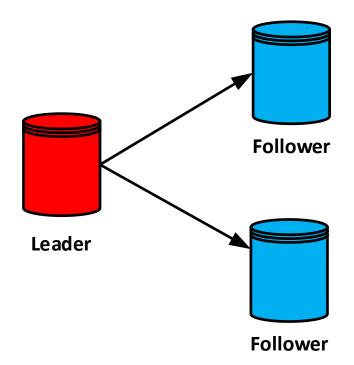
Репликация

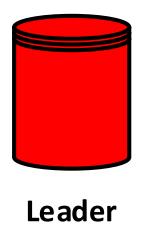
Репликация - поддержание актуальной копии данных на нескольких узлах.

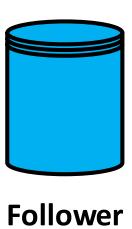
Репликационная группа или репликасет - множество узлов, между которыми настроена репликация данных

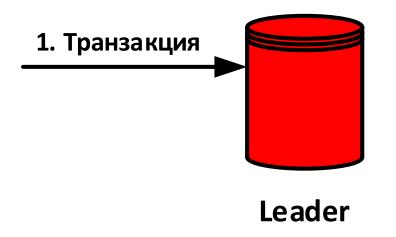
Задачи репликации:

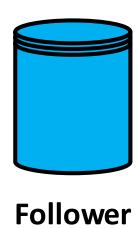
- Резервирование данных
- Балансировка нагрузки (обычно на чтение)

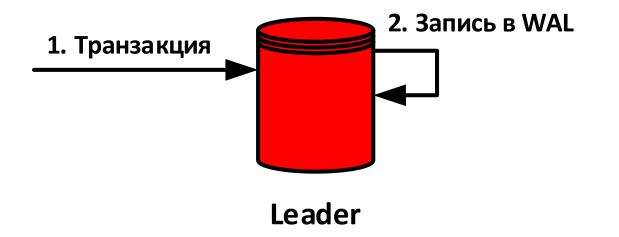


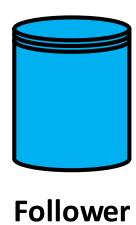


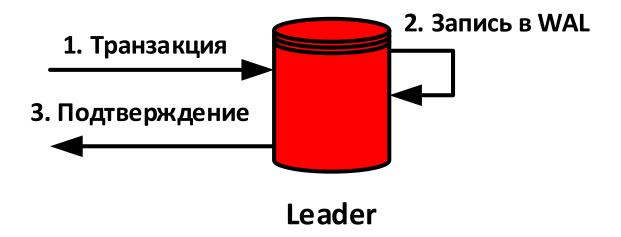


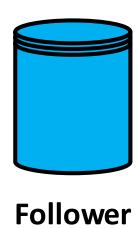


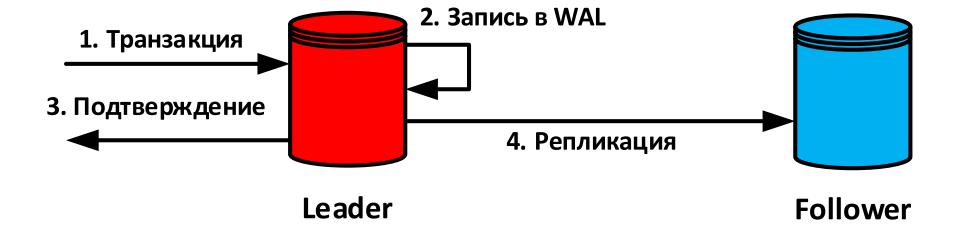














Асинхронная репликация. Плюсы и минусы

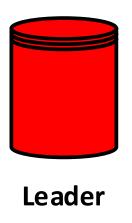


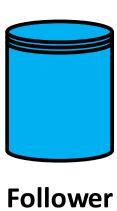
Преимущества

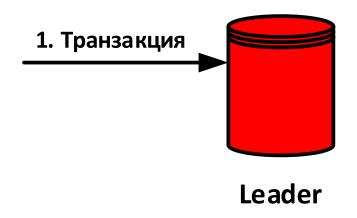
- Быстрый ответ, не ждём подтверждения от реплик
- Высокая доступность
- Легкая конфигурация

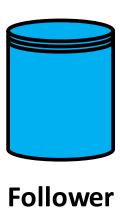
Недостатки

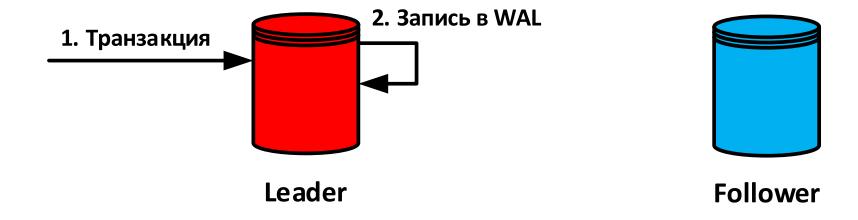
- > Возможность потери данных. Например, Leader может умереть после шага 3.
- Возможность грязных чтений.

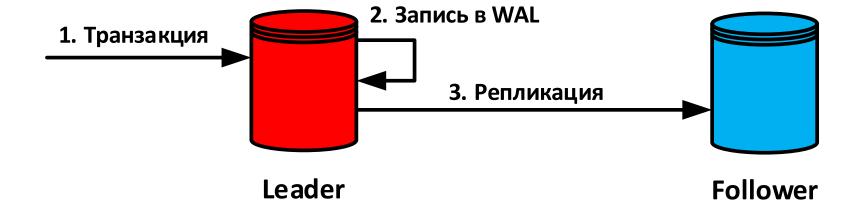


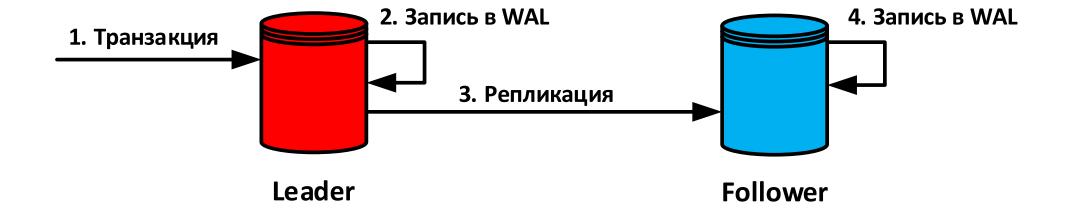


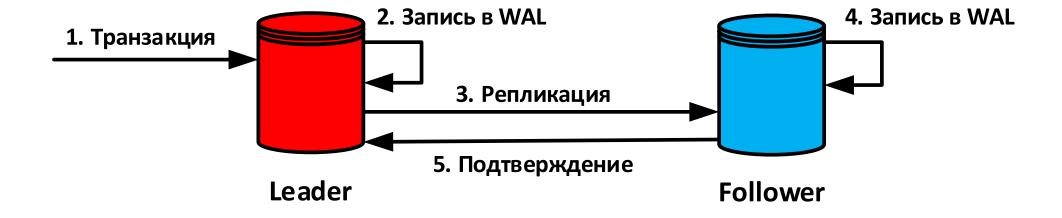




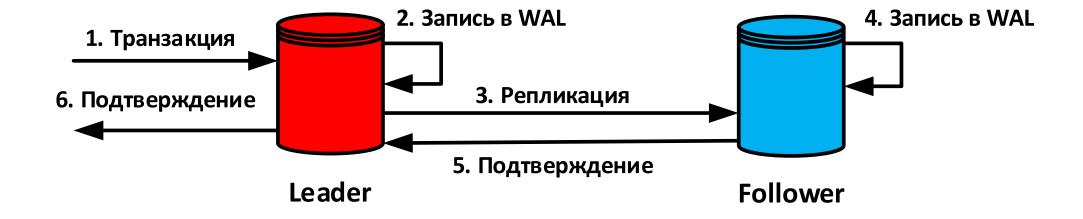












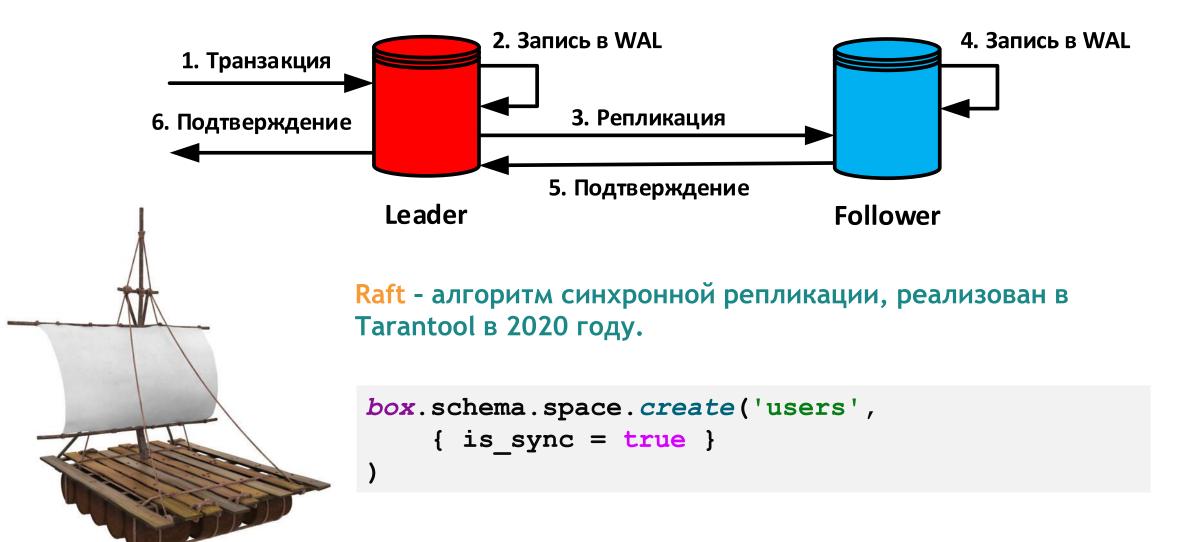
Синхронная репликация. Плюсы и минусы



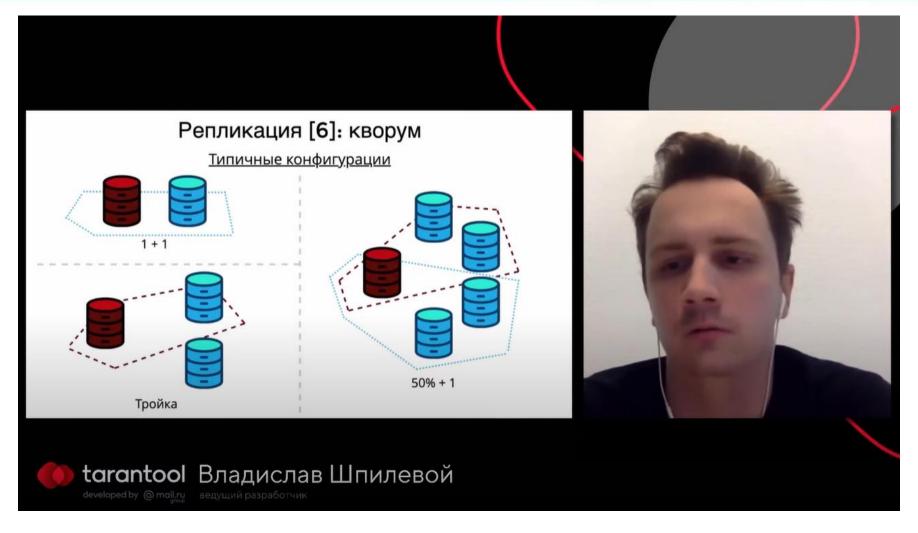
Преимущества

- Высокая надёжность. Данные сложнее потерять, чем при асинхронной репликации
 Недостатки
- > Низкая скорость (есть «тяжелые» походы по сети; ждём подтверждения от реплик)
- Низкая доступность (лидер может не набрать кворума для коммита транзакции)
- Сложность конфигурирования

Синхронная репликация в Tarantool. Алгоритм Raft



Tarantool 2.6 - Изобретая синхронную репликацию



https://www.youtube.com/watch?v=sOZgUQyPM_g

Шардирование данных в Tarantool

Шардирование - масштабирование БД по данным, данные распределяются по узлам кластера. В Tarantool используется шардирование по хешам, когда от каждой записи считается хеш, и записи с одинаковым хешем сохраняются на одних и тех же узлах кластера.

Основные реализации шардинга в Tarantool:

- Tarantool Shard https://github.com/tarantool/shard
 - Шардирование по хешам от первичных ключей записей
 - **Недостатки:** сложно хранить связанные данные вместе; медленный решардинг; нестабильные чтения при отказах нескольких реплик в репликасете
- Tarantool Vshard https://github.com/tarantool/vshard
 - Используются виртуальные хранилища (bucket) поверх реальных.
 - Bucket'ы переносятся между узлами атомарно.
 - Реализуется локальность данных по bucket_id.

Шардирование данных в Tarantool - Vshard

Шардирование данных в Tarantool - Vshard

local users = box.schema.space.create('users',

```
{ if not exists = true }
users:format({
    { 'bucket id', 'unsigned' },
    { 'uuid', 'string' },
    { 'login', 'string' },
    { 'password', 'string' },
    { 'status', 'string' }
})
            function create user (user uuid, login, password hash, status, groups)
                local bucket id = vshard.router.bucket id(login);
                local , err = vshard.router.callrw(bucket id, 'box.space.users:insert', {
                    { bucket id, user uuid, login, password hash, status }
                })
                insert user groups(bucket id, user uuid, groups)
                return user uuid
            end
```

Шардирование данных в Tarantool - Vshard

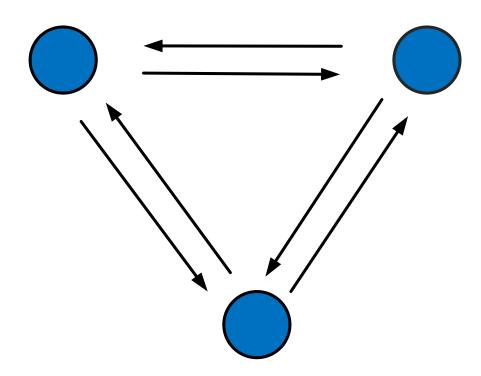


Требования

Необходимо уметь обнаруживать отказы узлов (failure detection) в кластере для возможности принятия согласованных решений

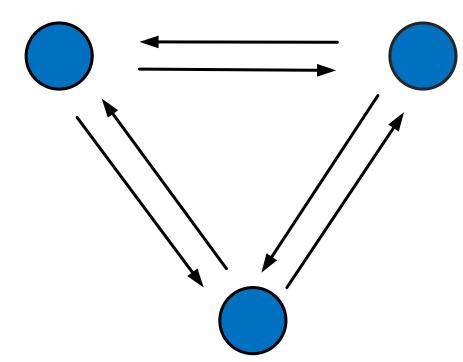
Требования

Необходимо уметь обнаруживать отказы узлов (failure detection) в кластере для возможности принятия согласованных решений



Требования

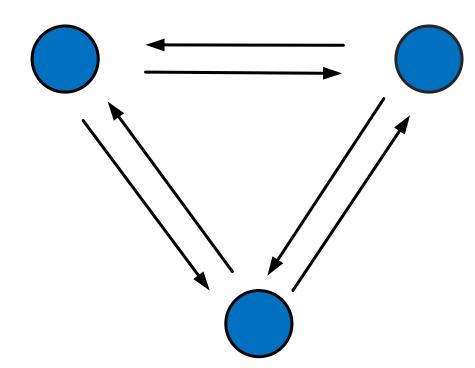
Необходимо уметь обнаруживать отказы узлов (failure detection) в кластере для возможности принятия согласованных решений



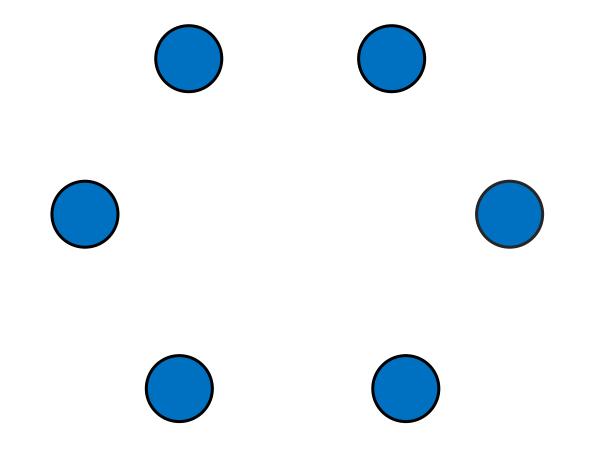
При N узлах будет $O(N^2)$ сообщений в сети. Это не работает на больших кластерах.

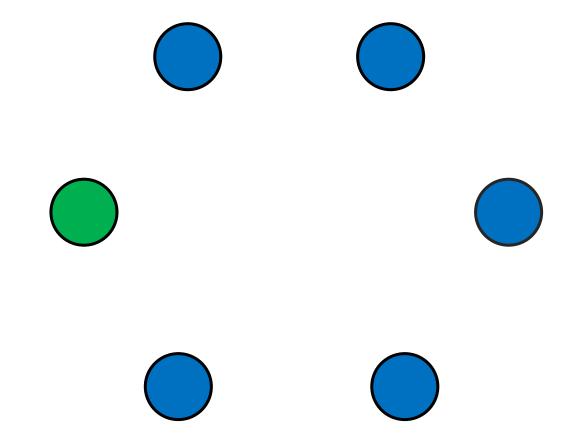
Требования

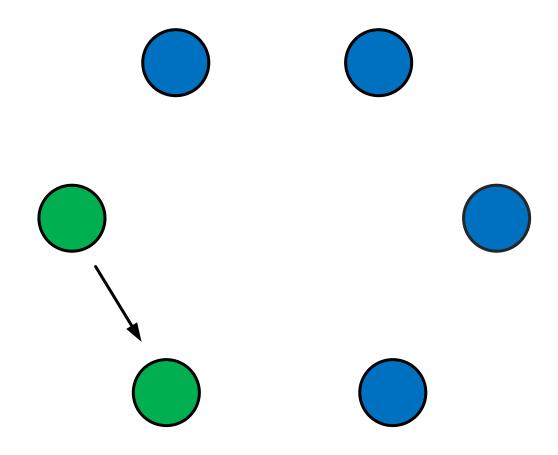
- Необходимо уметь обнаруживать отказы узлов (failure detection) в кластере для возможности принятия согласованных решений
- Необходимо делать это эффективно

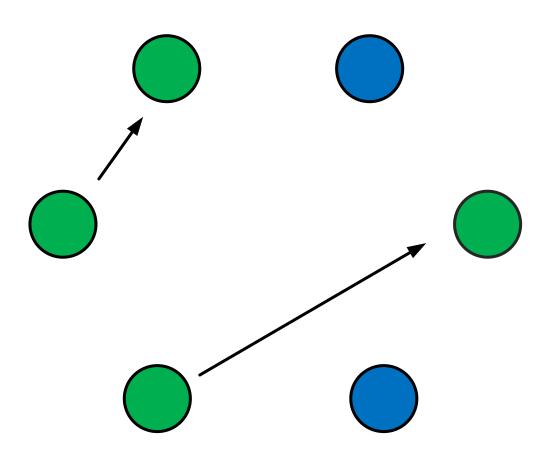


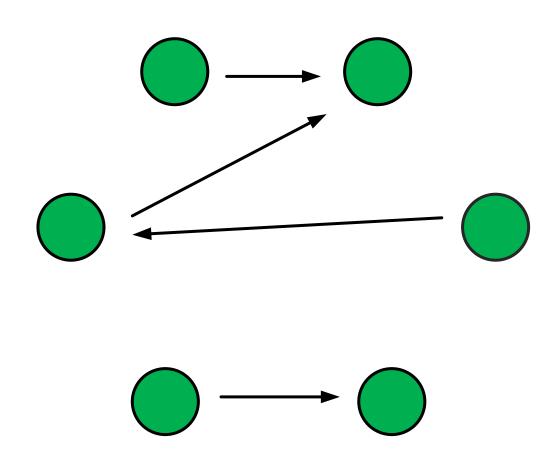
При N узлах будет $O(N^2)$ сообщений в сети. Это не работает на больших кластерах.

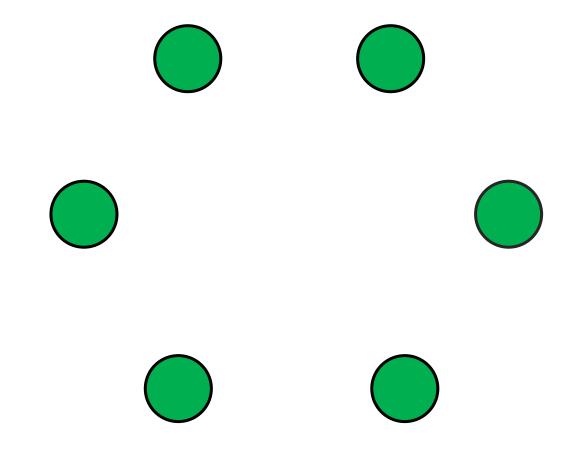






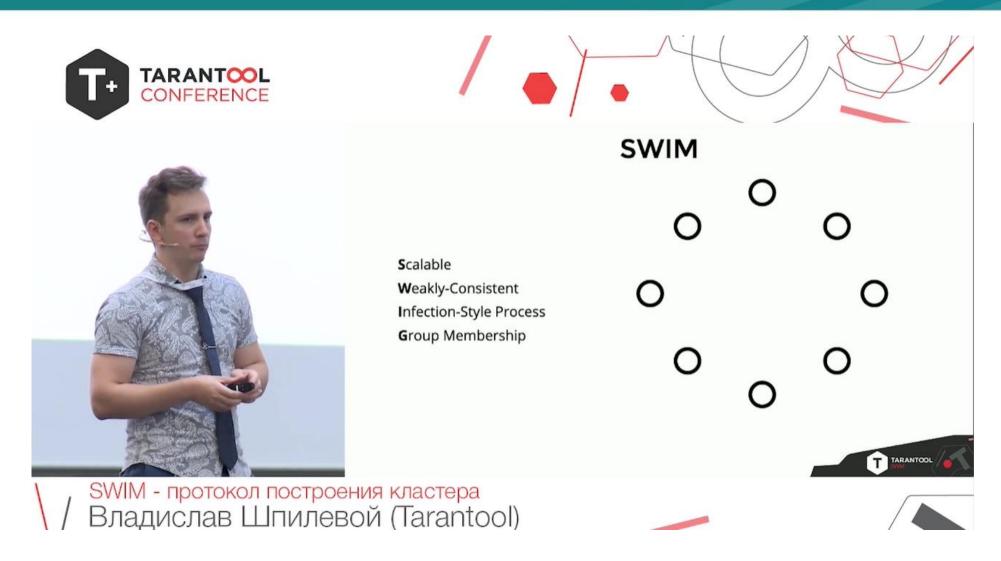








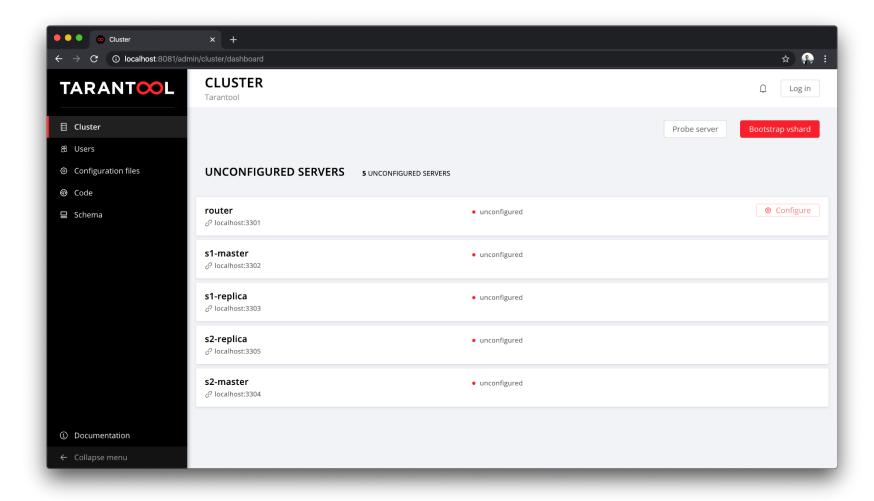
SWIM B Tarantool



Tarantool Cartridge

cartridge create --name myapp

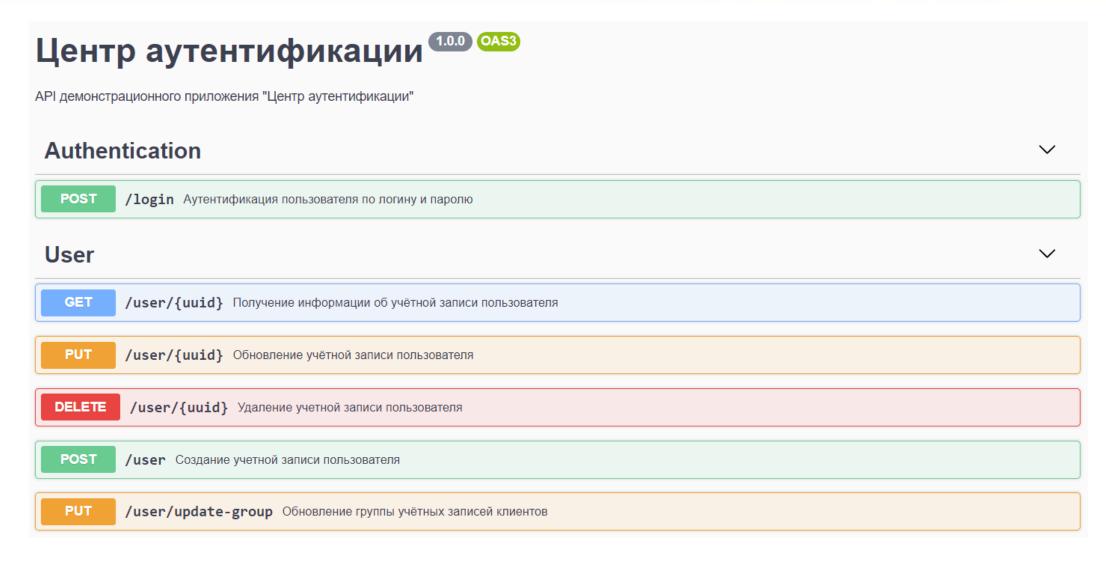
cd myapp
cartridge build
cartridge start



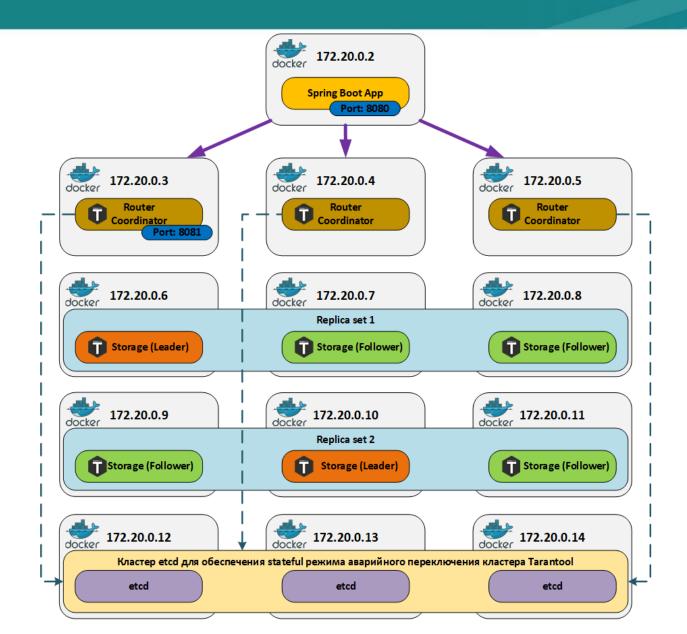
Tarantool сегодня

Charding		Tarantool VShard
Sharding		Tarantooi vanaru
Replication	· ·	Synchronous/Asynchronous
In-memory		memtx engine
Disk	9	vinyl engine, LSM-tree
Persistency		Both engines
SQL	0	ANSI
Stored procedures	9	Lua, C, SQL
Audit logging	0	Yes
Connectors to DBMSes	•	MySQL, Oracle, Memcached
Static build	0	for Linux
GUI	0	Cluster management
Unprecedented performance	0	100.000 RPS per instance - easy

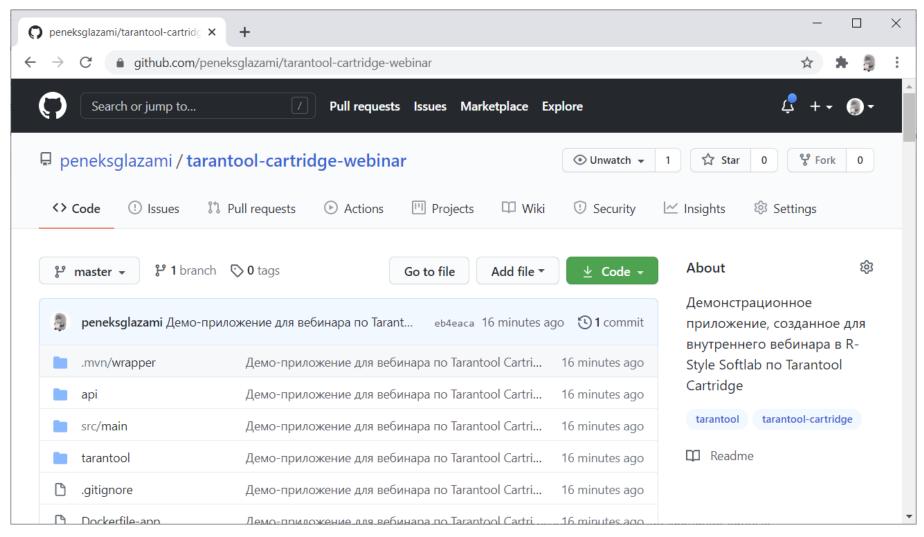
API демо-приложения



Демо-приложение. Разбиение приложений по контейнерам



Время переходить в IDE



https://github.com/peneksglazami/tarantool-cartridge-webinar

Как посмотреть демо-приложение в Gitpod

Запуск демонстративного кластера в Gitpod

Для запуска демо-приложения в облаке gitpod.io необходимо выполнить следующие действия:

- 1. Создайте учётную запись на gitpod.io.
- 2. Необходимо в личном кабинете gitpod.io зайти в раздел Settings и проставить галочку напротив настойки "Enable Feature Preview". Это позволит запускать docker-контейнеры внутри docker-контейнера с облачной IDE.
- 3. Запустить workspace с приложением. Создать workspace можно, перейдя по ссылке

https://gitpod.io/#https://github.com/peneksglazami/tarantool-cartridge-webinar.

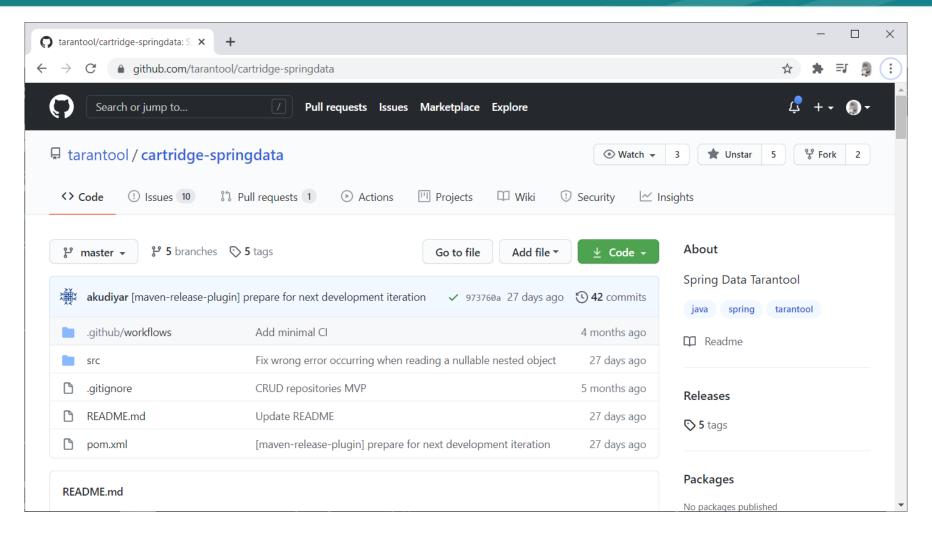




- 4. В терминале выполните команду sudo docker-up . Эта команда запустит демон docker.
- 5. Откройте новое окно терминала и далее выполните команды запуска кластера, описанные выше в разделе "Запуск демонстративного кластера через docker-compose".

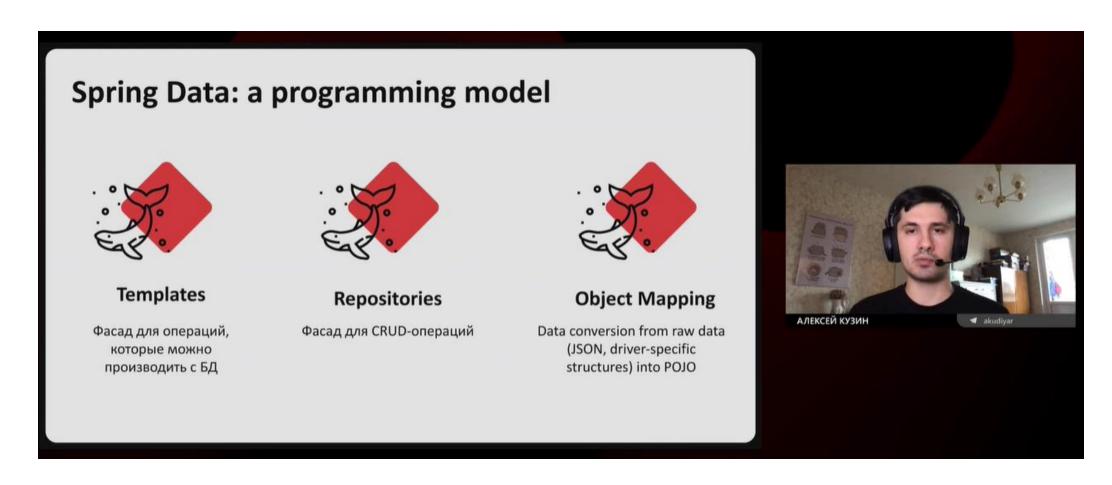
https://gitpod.io/#https://github.com/peneksglazami/tarantool-cartridge-webinar

Интеграция со Spring Data



https://github.com/tarantool/cartridge-springdata

Алексей Кузин — Работа с шардированными данными в памяти со вкусом Spring Data



https://www.youtube.com/watch?v=tjKK0O2XrFo

Tarantool Cartridge в Kubernetes



https://www.youtube.com/watch?v=8NvE6uooMQY

Конкурс

Заходим на menti.com

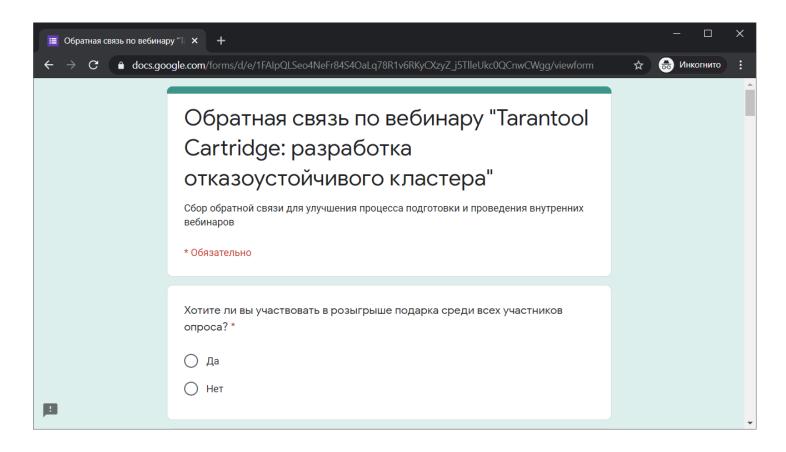
39 20 89 9



Подарок победителю!



Обратная связь





https://forms.gle/j1Nd5C8FKt3AUhVe9

Розыгрыш приза среди участников опроса





https://youtu.be/X2EIZsVva4c

Розыгрыш приза состоится 9 марта в 9:30