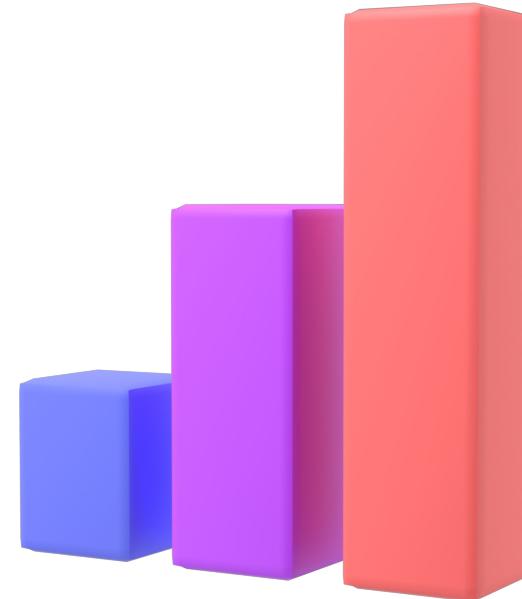
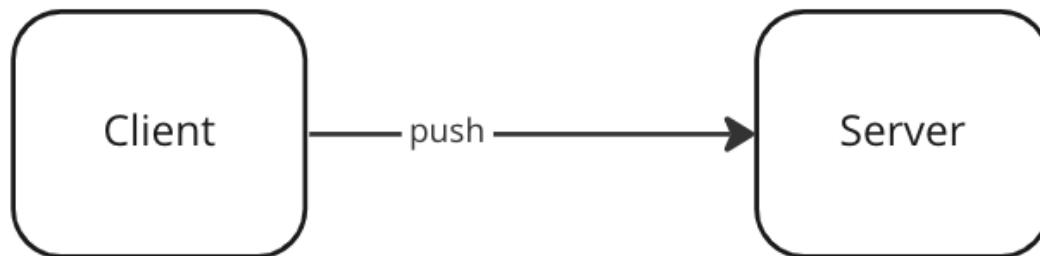


System Design

Паттерны и приемы проектирования



push / pull модель



Репликация

Бэкап это или нет?

Репликация и бэкапы

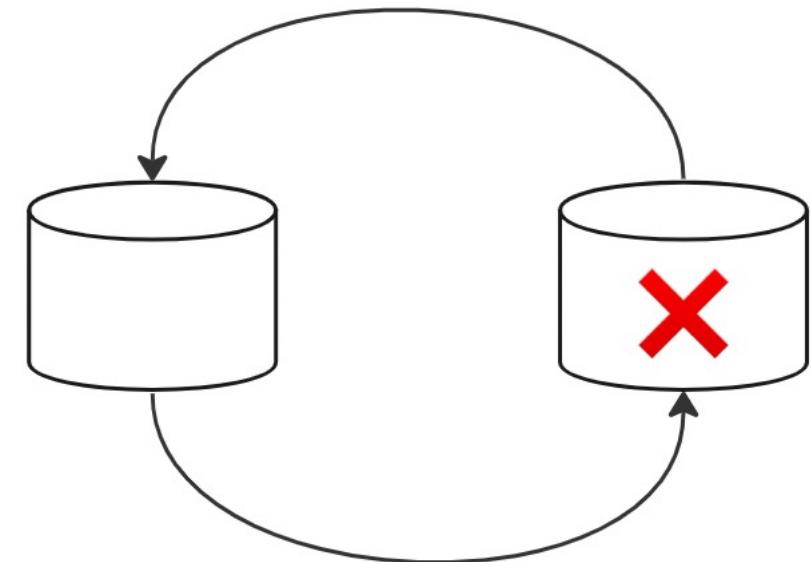
Бэкап – это **резервное копирование** содержимого диска с целью последующего восстановления.

Репликация – это создание клона базы данных для быстрого подхвата функций поврежденной системы



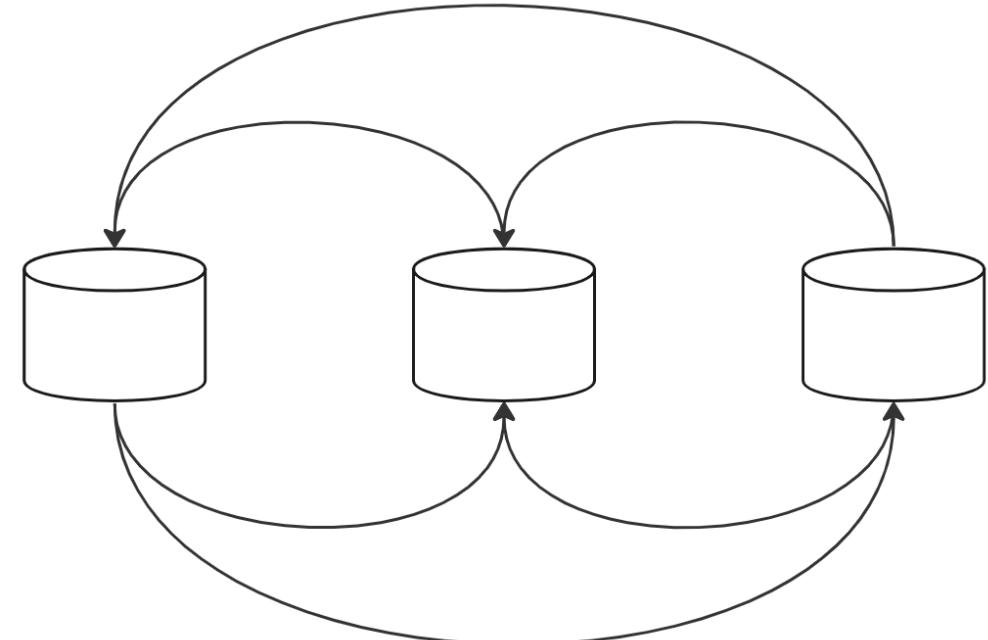
Надежность

Поддержка резервной базы данных на
случай потери основной



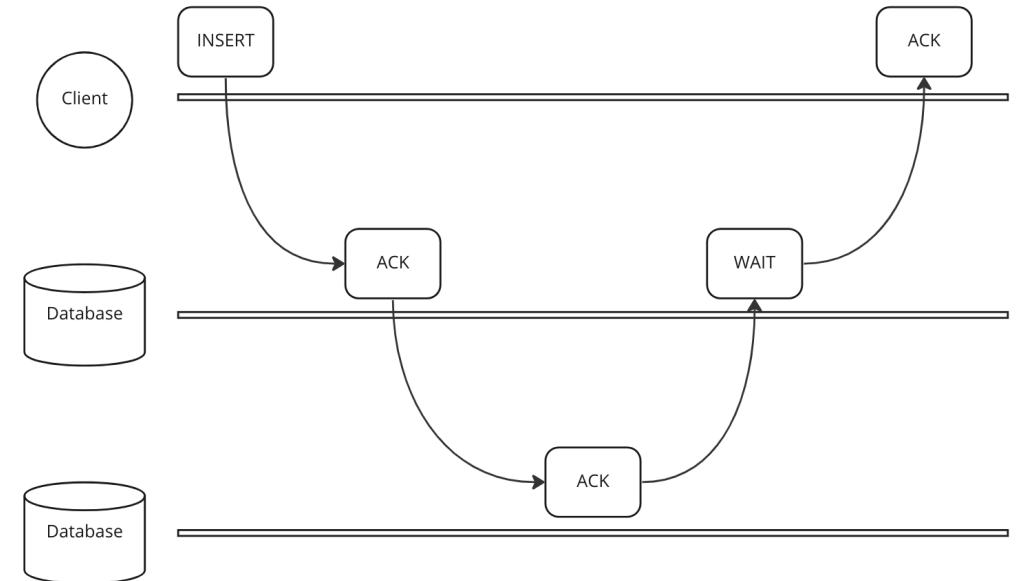
Масштабирование чтения

Снижение нагрузки на чтение за счет
переноса части запросов на реплики



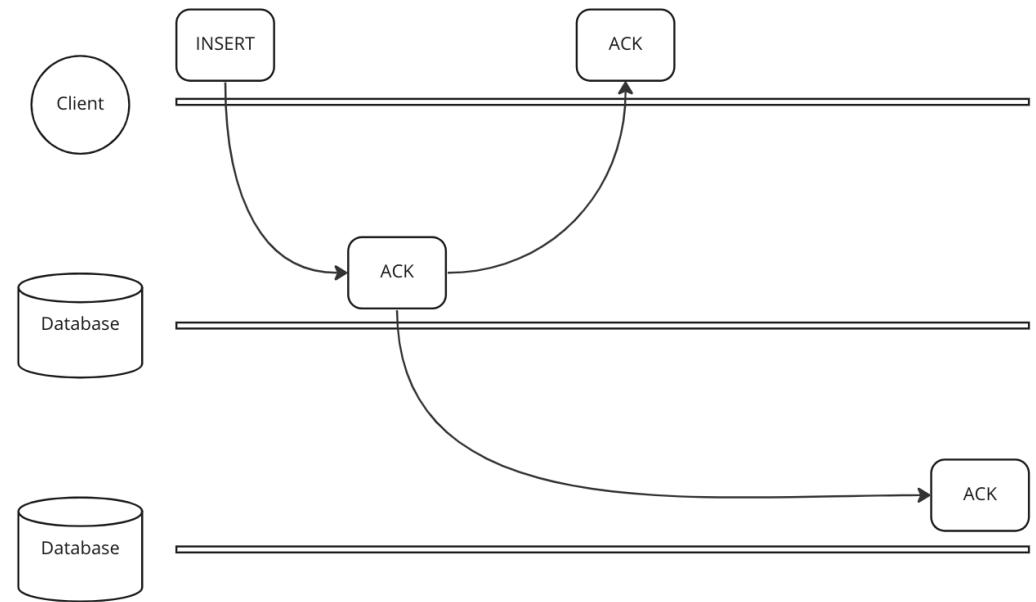
Синхронная

1. Запись транзакции в журнал
2. Применение транзакции в движке
3. Отправка данных на все реплики
4. Получение подтверждения от всех реплик
5. Возвращение подтверждения клиенту

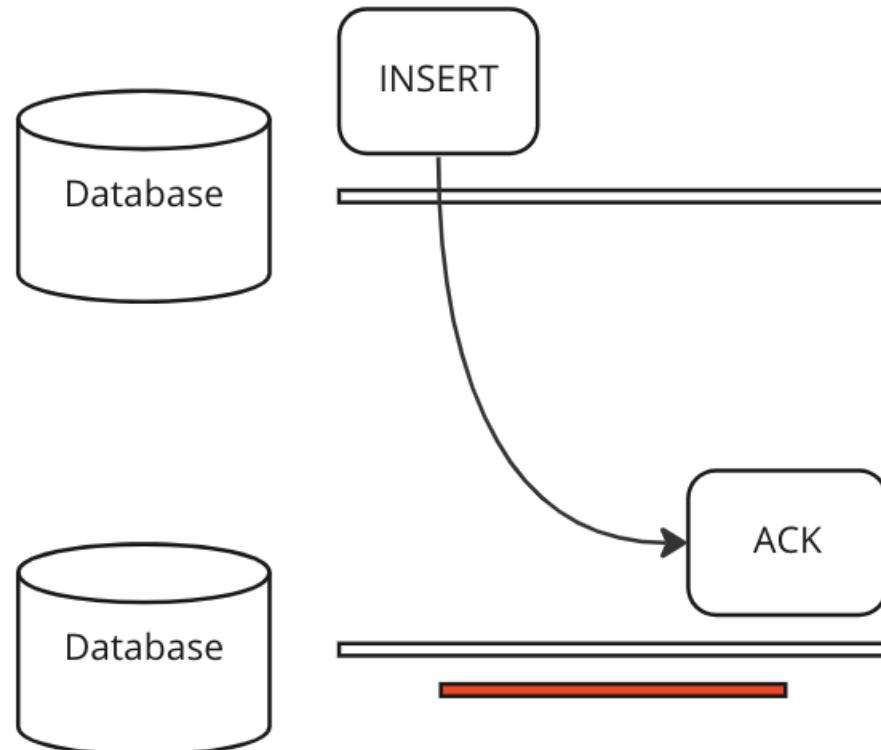


Асинхронная

1. Запись транзакции в журнал
2. Применение транзакции в движке
3. Возвращение подтверждения клиенту
4. Отправка данных на реплики

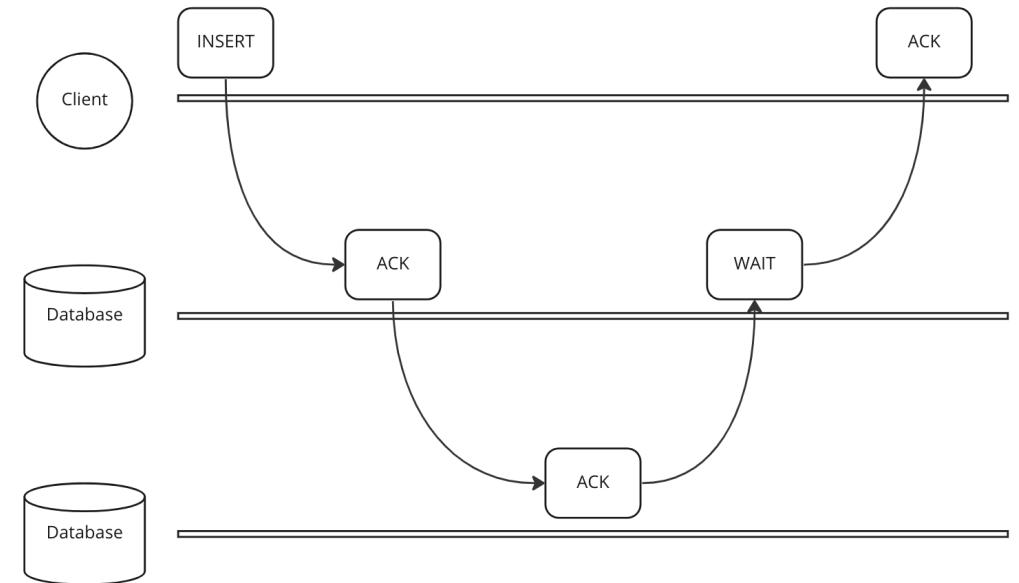


Replication Lag



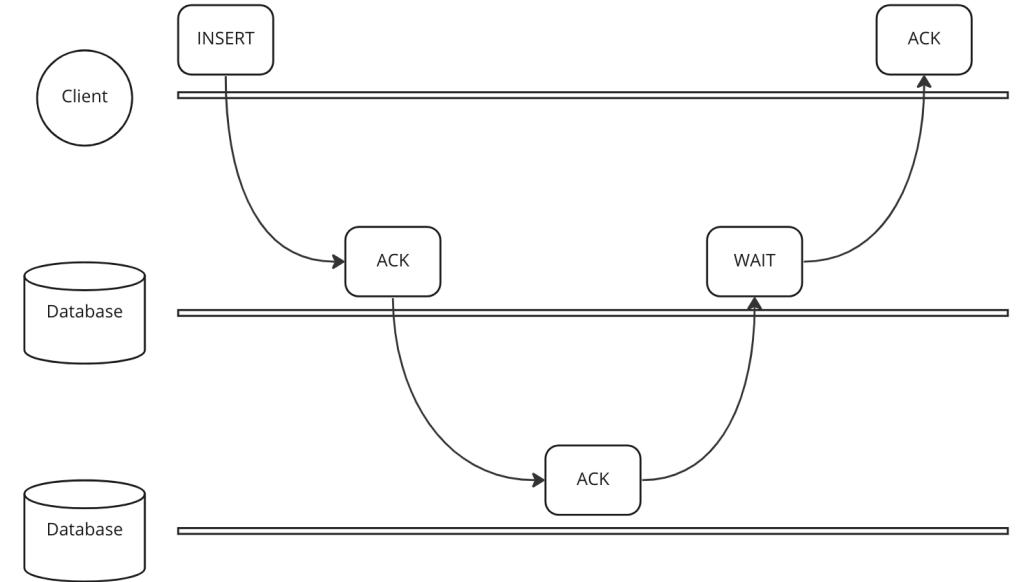
Полусинхронная (semisync)

1. Запись транзакции в журнал
2. Применение транзакции в движке
3. Отправка данных на реплики.
4. Получение подтверждения от реплики о получении изменений (применены они будут «когда-то потом»)
5. Возвращение подтверждения клиенту



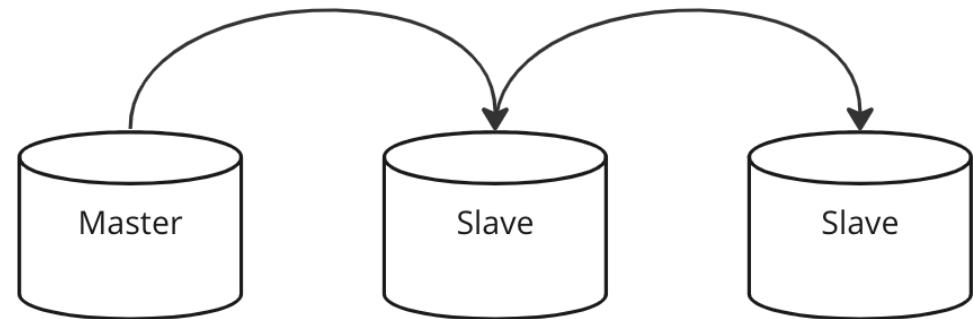
Lose - less semisync

1. Запись транзакции в журнал
2. Отправка данных на реплики.
3. Получение подтверждения от реплики о получении изменений (применены они будут «когда-то потом»)
4. Применение транзакции в движке
5. Возвращение подтверждения клиенту

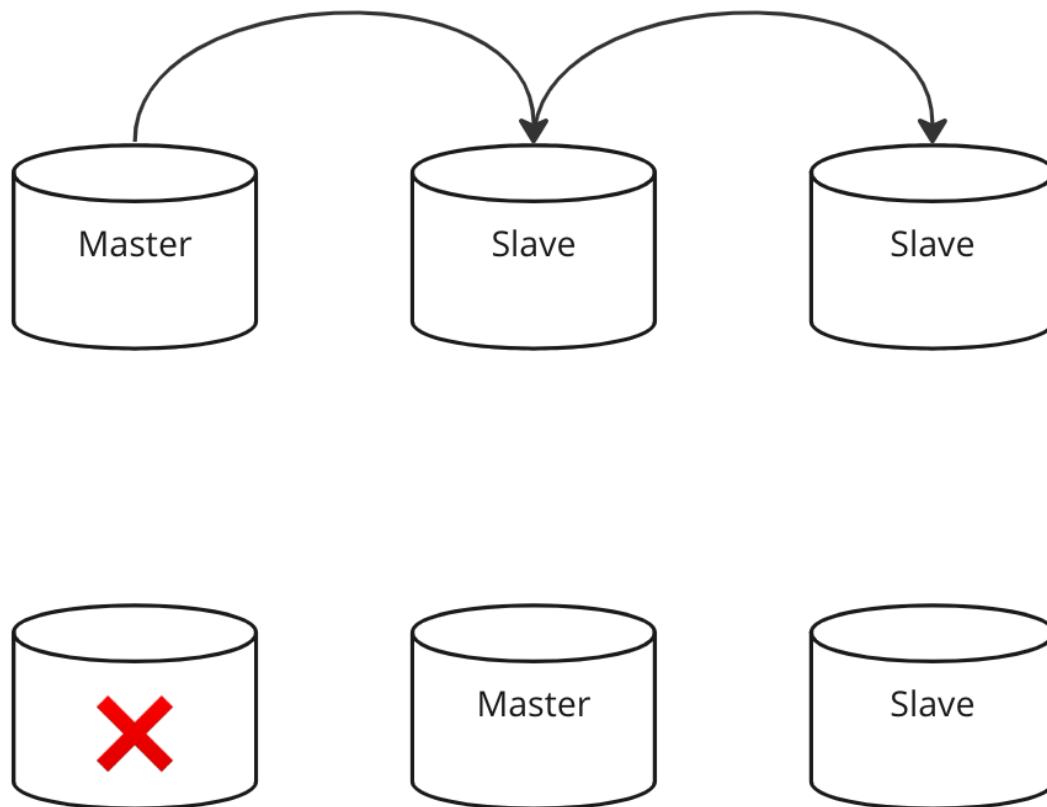


Master - slave

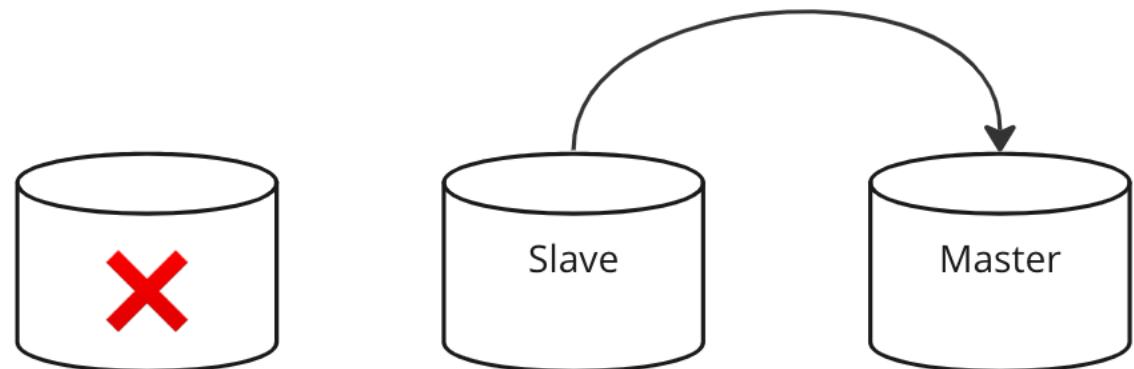
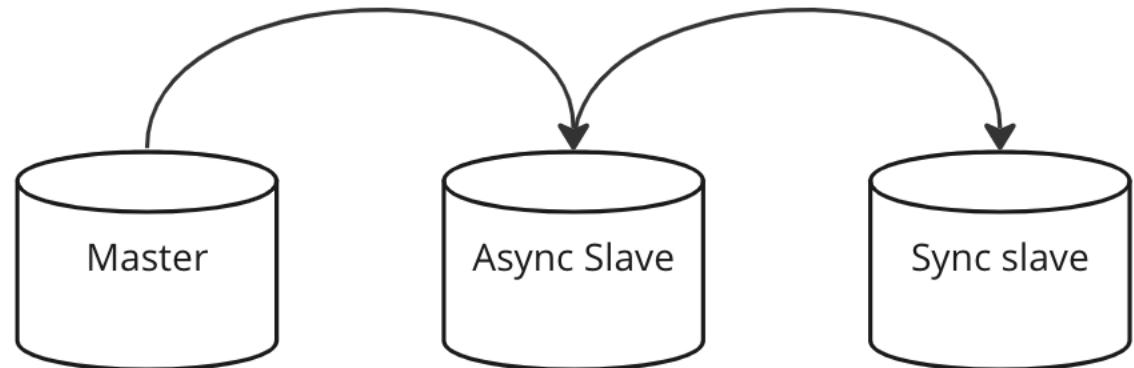
- Пишем в мастер
- Читаем из слейвов или из мастера
- В случае падения мастера получаем downtime на запись



Failover



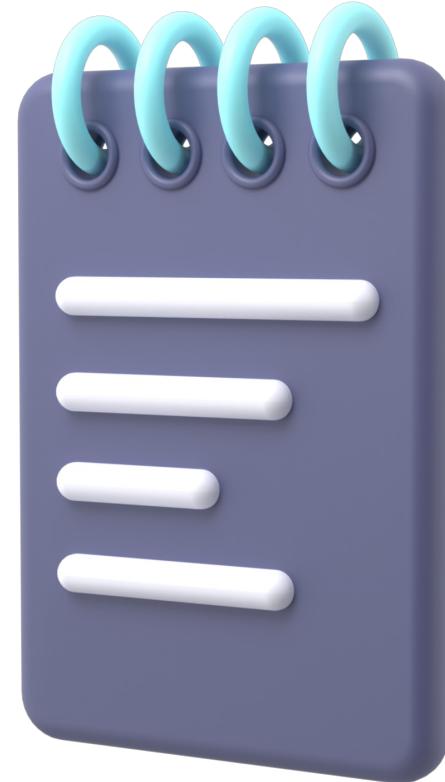
Hot Standby



Допустим, вы создаете клон Twitter.

Пользователь может опубликовать твит со своего компьютера, который отправит запрос на запись в вашу распределенную базу данных.

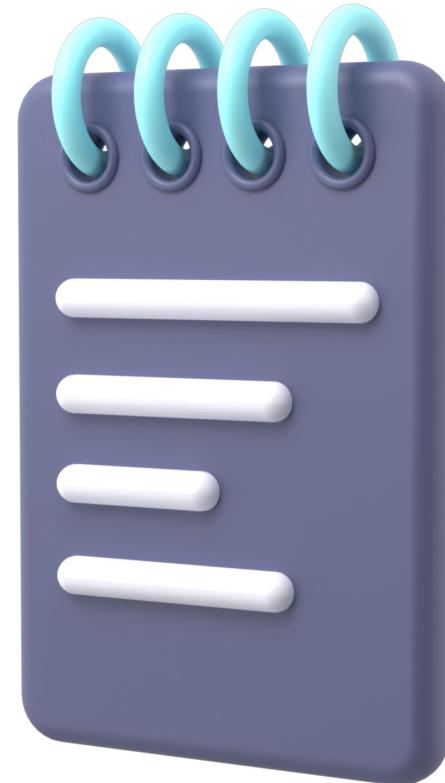
Этот запрос на запись реплицируется асинхронно, поэтому лидер ответит, что запись прошла успешно после изменения его локального состояния. Затем он отправит изменение всем узлам-последователям.



Чтение собственных записей

Нужно отследить, когда пользователь в последний раз отправлял обновление. Если он отправил обновление в течение последней минуты, то запросы на чтение должен обрабатываться главным узлом

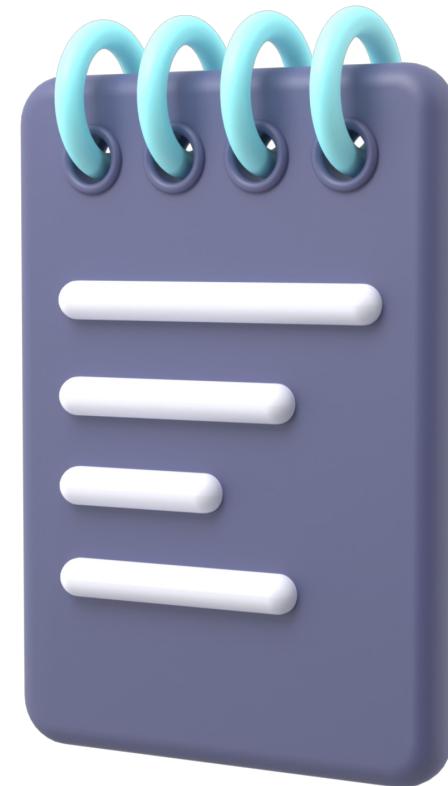
Возвращаясь к примеру с клоном Twitter,
асинхронная репликация приведет к тому,
что некоторые базы-фоловеры будут
отставать от других узлов с точки зрения
обновлений.



Монотонное чтение

Обеспечить, чтобы каждый пользователь всегда читал из одного и того же узла-последователя (разные пользователи могут читать с разных реплик)

Допустим, у вас есть пользователь А и пользователь Б в вашем приложении-клоне Twitter. Пользователь А публикует в Твиттере фотографию своей собаки. Пользователь Б отвечает на это фото в твиттере комплиментом собаке.



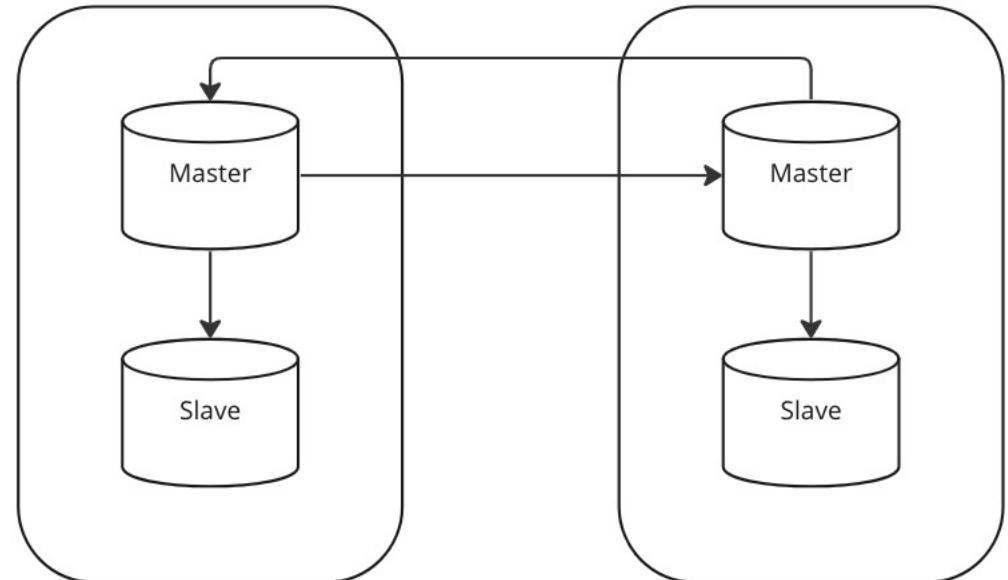
Существует причинно-следственная связь между двумя твитами, когда ответный твит пользователя Б не имеет никакого смысла, если вы не видите твит пользователя А.

Согласованное префиксное чтение

Гарантировать, что база данных всегда применяет операции в одном и том же порядке, а именно писать в одну и туже секцию

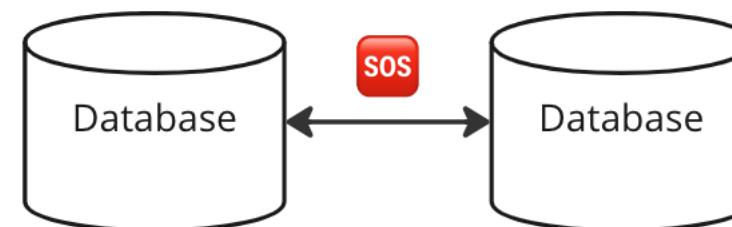
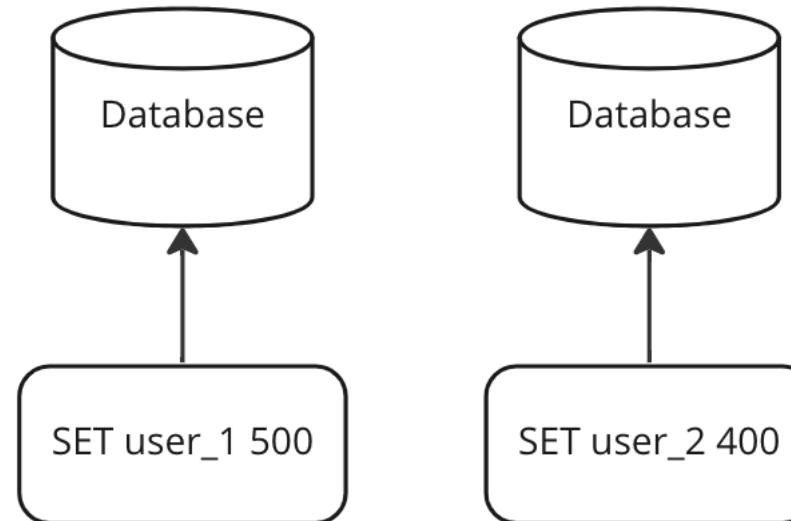
Master - master

- Появляются конфликты
- Пишем в разные мастера
- Читаем из слейвов или мастеров
- В случае падения мастера нет никакого downtime на запись



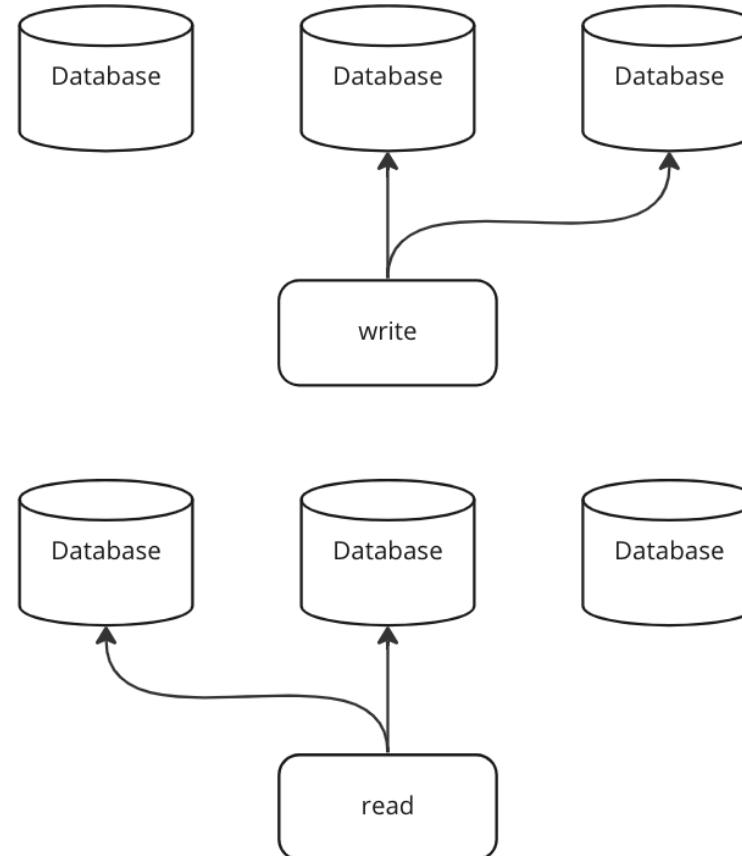
Конфликты

- LWW
- Ранг реплик
- Решение конфликтов на клиенте
- Conflict-replicated data type (CRDT)



Master - less

- Пишем во все ноды
- Читаем со всех нод
- Запись или чтение считается успешными по формуле $(w + r > n)$



Как быть с консистентностью?

Обновления при чтении или противодействие энтропии

Источник передачи данных

1. **push** – мастер рассыпает данные репликам (PgSQL)
2. **pull** – реплики стягивают данные сами (MySQL)

Форматы передачи данных

Что будем и как будем передавать?

Statement base

1. Передаются запросы (но не сущности)
2. Каждый запрос считается на каждой ноде (random(), unix_timestamp(), ...)

Row based

1. Передаются измененные строчки в бинарном виде
2. Передаются сущности, даже если изменили одно поле
(но можно настраивать full / minimal, ...)

Mixed

База данных переключается из SBR в RBR или из
RBR в SBR, в зависимости от той ли иной ситуации

Логическая репликация

- Работает с кортежами (SBR, RBR)
- Не знает, как они хранятся на диске

Физическая репликация

- Работает со страницами
- slave = master (байт в байт)

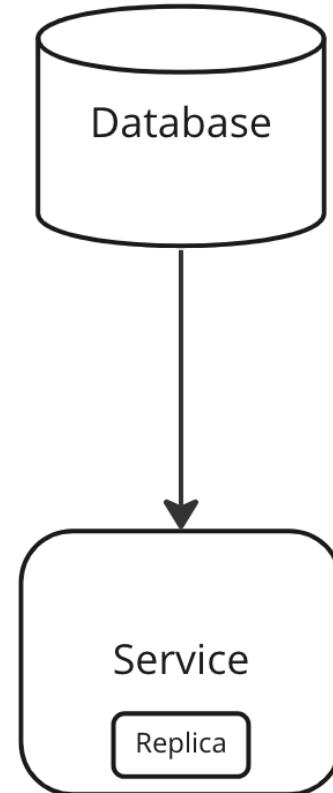
libslave

ABOUT

This is a library that allows any arbitrary C++ application to connect to a Mysql replication master and read/parse the replication binary logs.

In effect, any application can now act like a Mysql replication slave, without having to compile or link with any Mysql server code.

One important use-case for this library is for receiving changes in the master database in real-time, without having to store the master's data on the client server.



Фильтрация репликаций

Можно реплицировать данные частично

CAP теорема

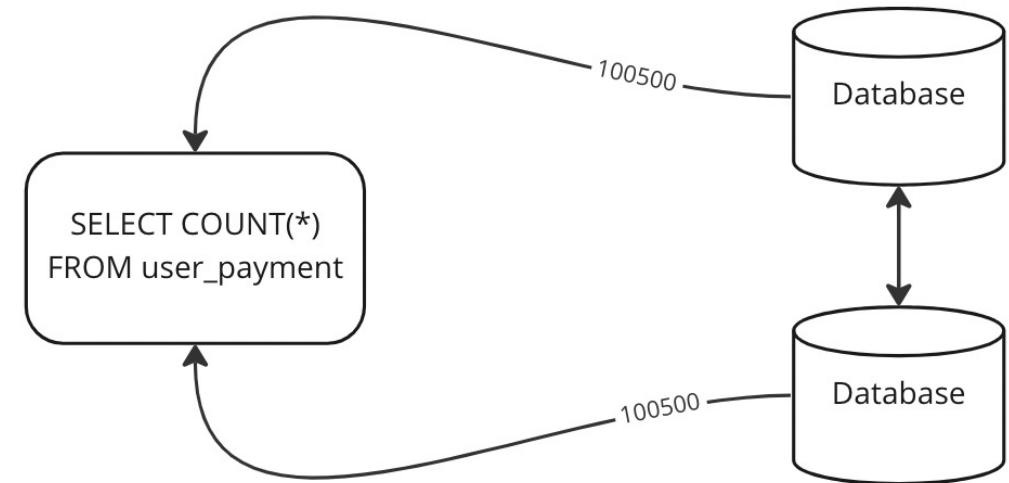
О чем она гласит?

CAP теорема

Утверждение о том, что в любой реализации распределенных вычислений возможно обеспечить не более двух из трёх следующих свойств (согласованность, доступность, устойчивость к разделению)

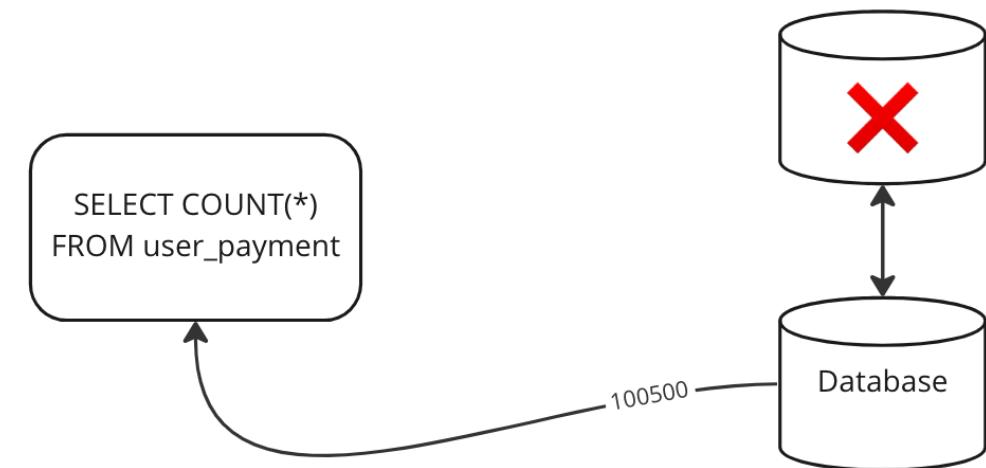
Консистентность

Данные во всех нодах одинаковы



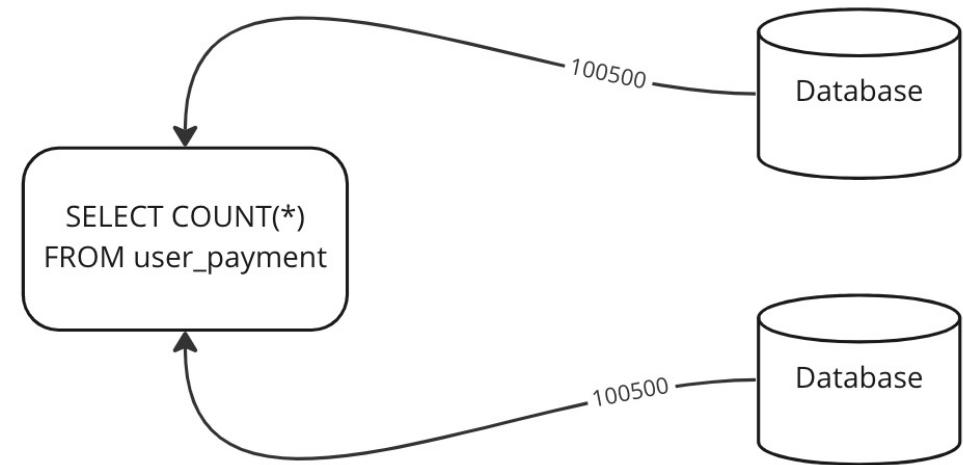
Доступность

Если запрос пришел на живую ноду,
запрос будет получен за конечное время



Устойчивость к разделению

Продолжаем работать не смотря
на отсутствие связи



Обрубаем связь

1. Обрубаем все запросы – **AC система**
2. Разрешаем из нод читать, но запрещаем читать – **CP система**
3. Разрешаем читать и писать – **AP система**

Партиционирование

Метод разделения больших таблиц на много маленьких,
и желательно, чтобы это происходило прозрачным для
приложения способом (**секции находятся на одном и**
том же инстансе базы данных)



Вертикальное

ID	Name	Status	Description	Photo
10				
20				
30				
50				

ID	Name	Status
10		
20		
30		
50		

ID	Description	Photo
10		
20		
30		
50		

Горизонтальное

ID	Name	Status	Description	Photo
10				
20				
30				
50				

ID	Name	Status	Description	Photo
10				
20				

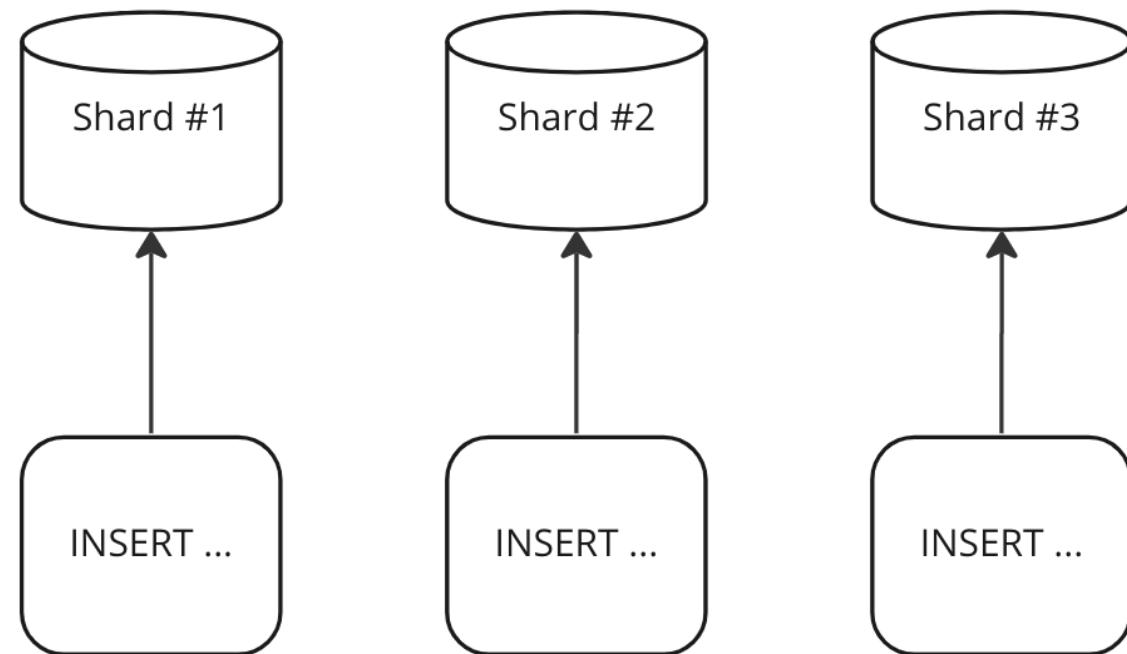
ID	Name	Status	Description	Photo
30				
50				

Шардирование

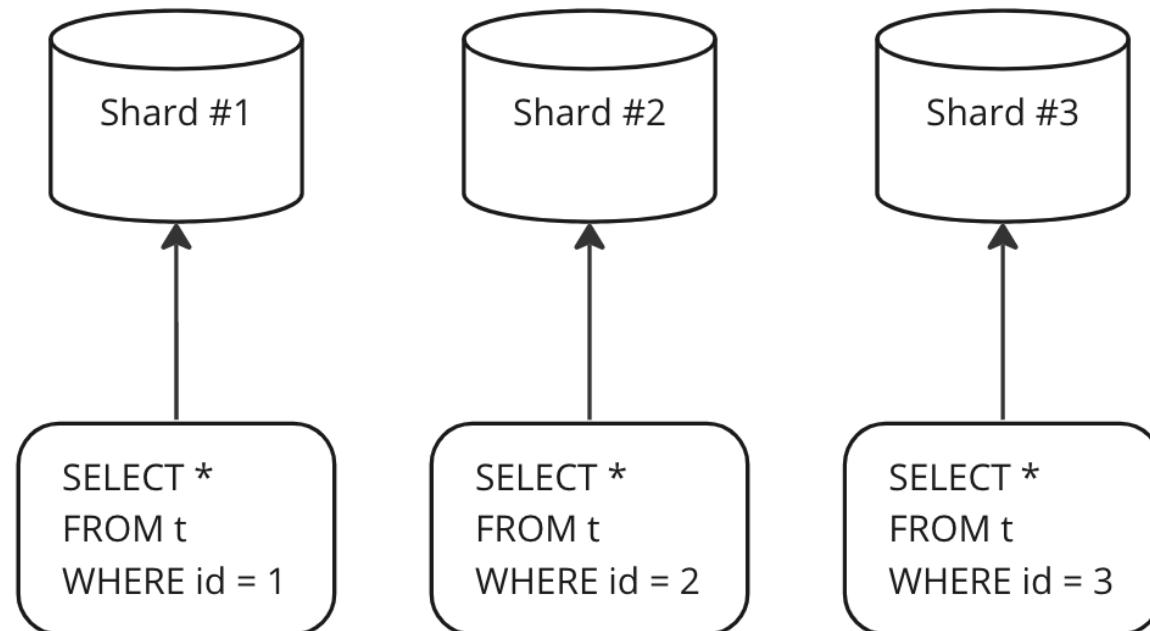
Подход, предполагающий разделение таблиц на независимые сегменты, каждый из которых управляет отдельным инстансом базы данных



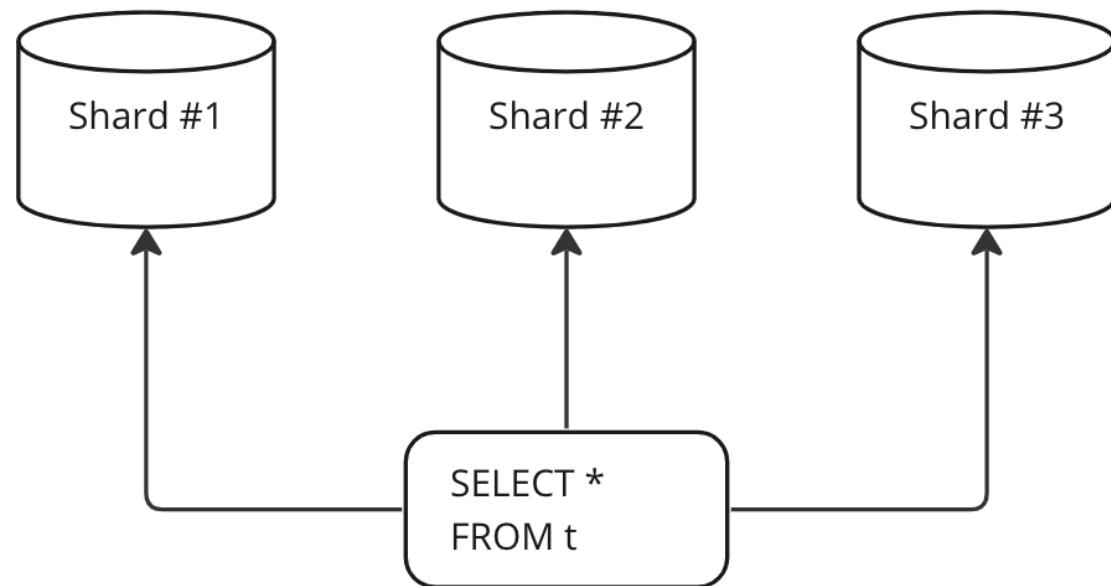
Шардирование



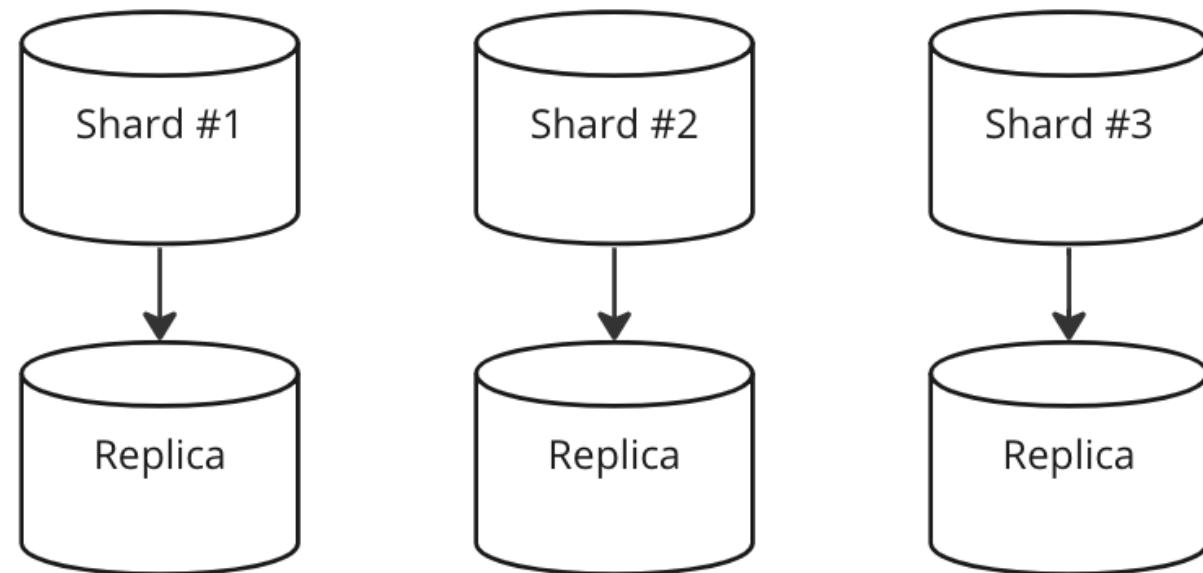
Шардирование



Шардирование



Шардирование



Range based

ID	Price
10	30
20	57
30	64
50	123

Shard #1 (0...50)

10	30
----	----

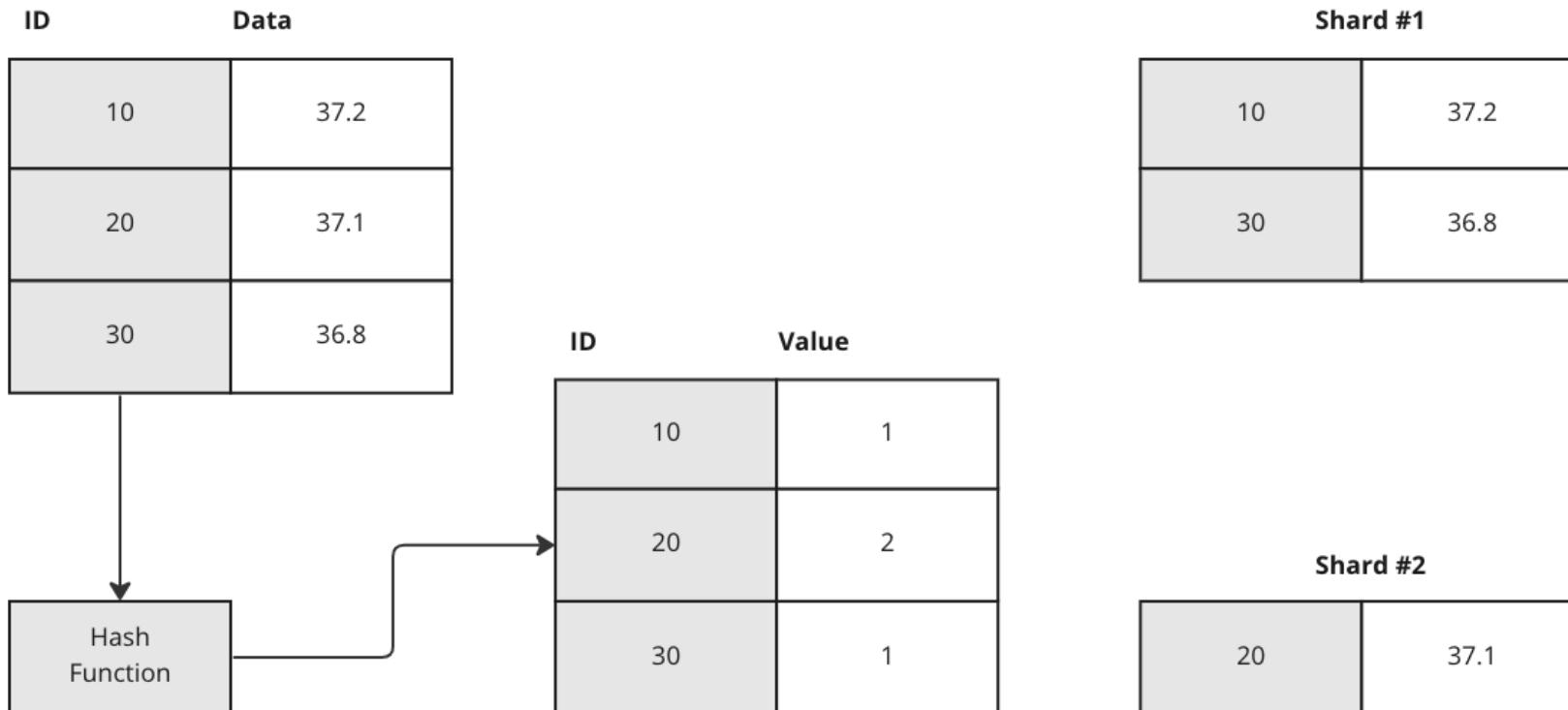
Shard #2 (50...100)

20	57
30	64

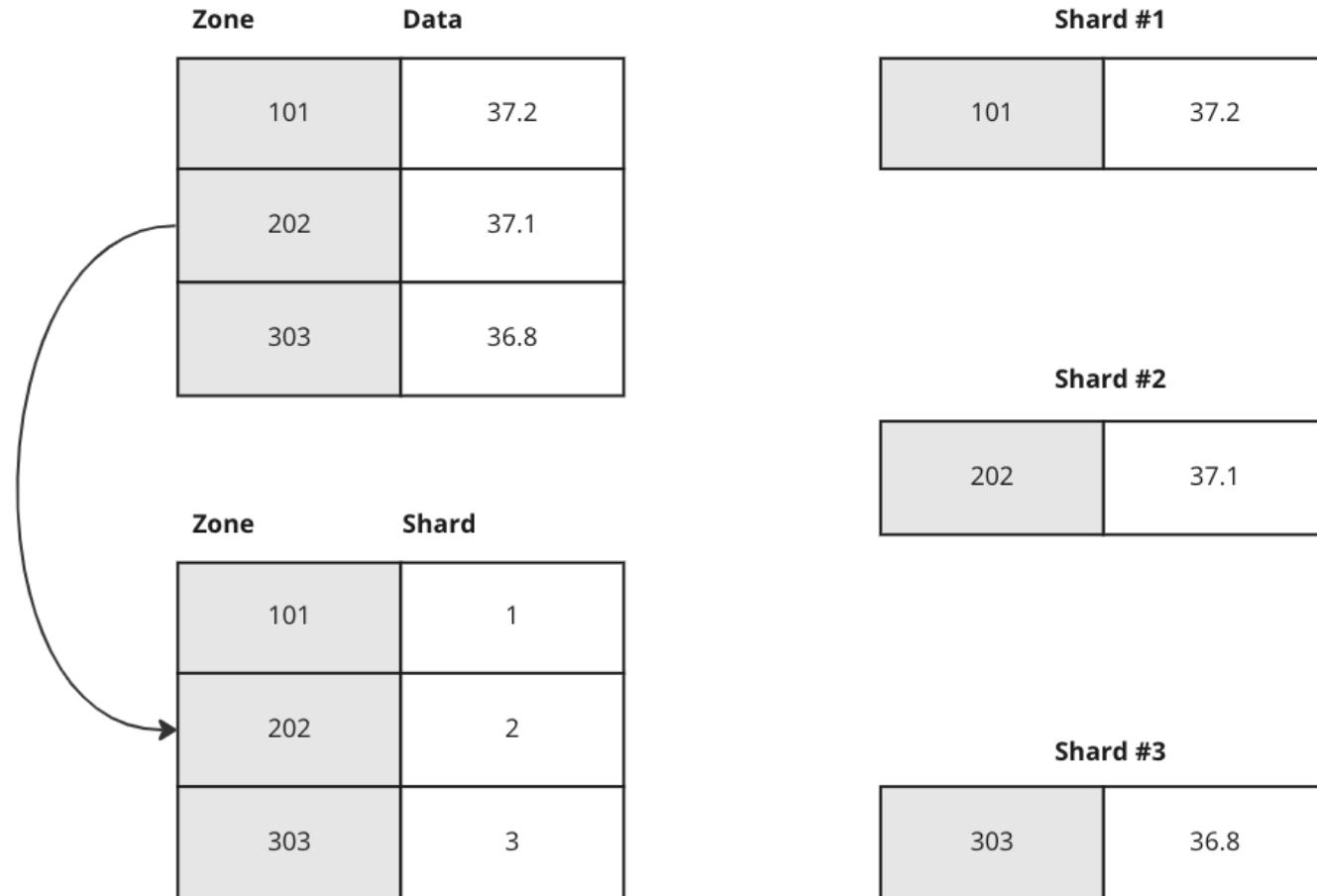
Shard #3 (100+)

50	123
----	-----

Key based



Directory based

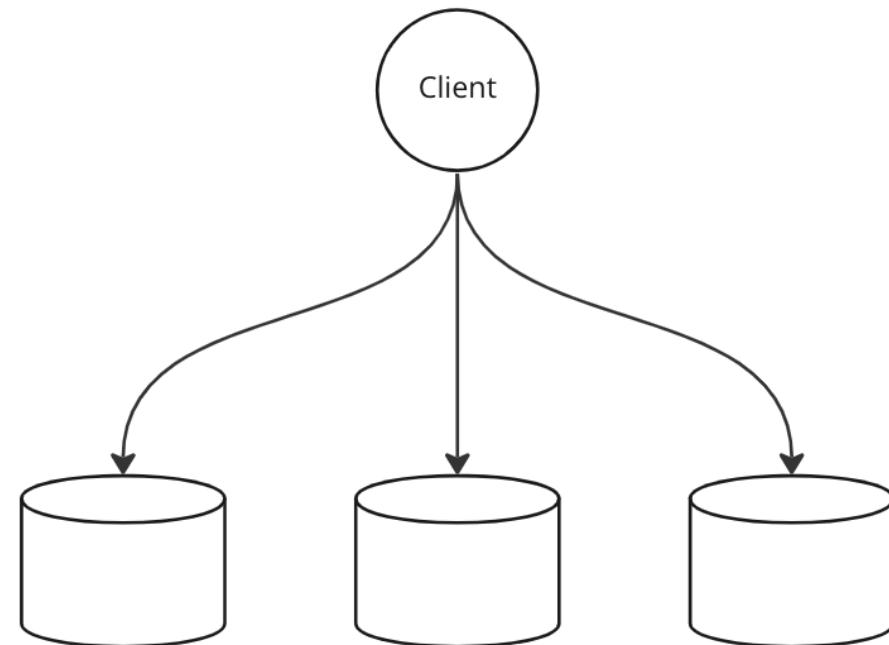


Routing

Как понять куда идти?

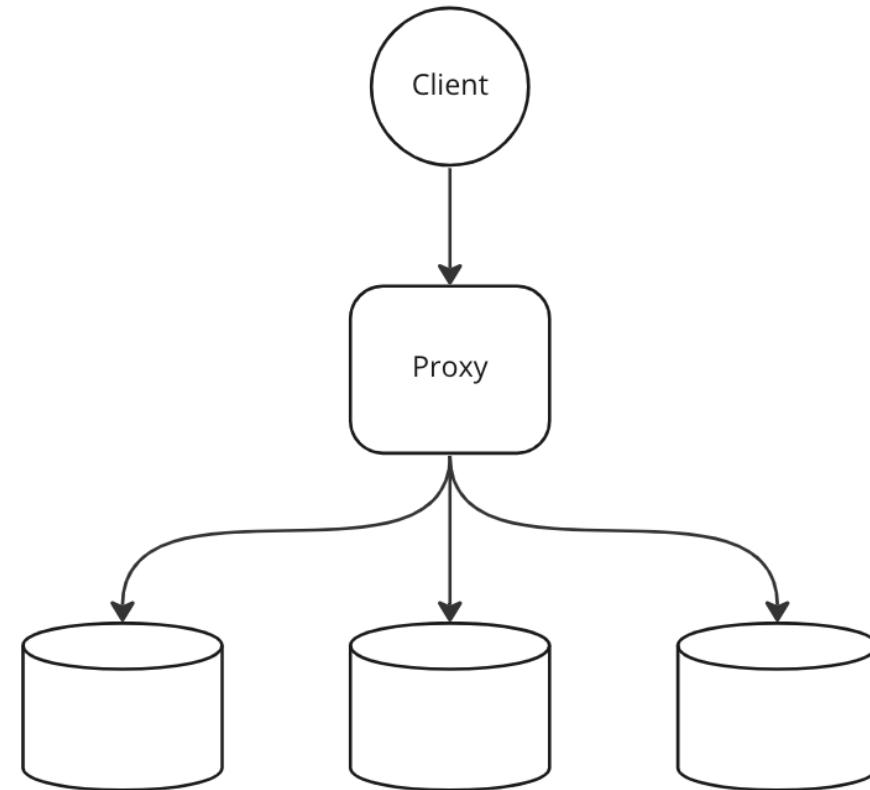
Клиентский

- + нет лишних узлов
- дополнительная логика в клиенте
- сложности с обновлением хостов



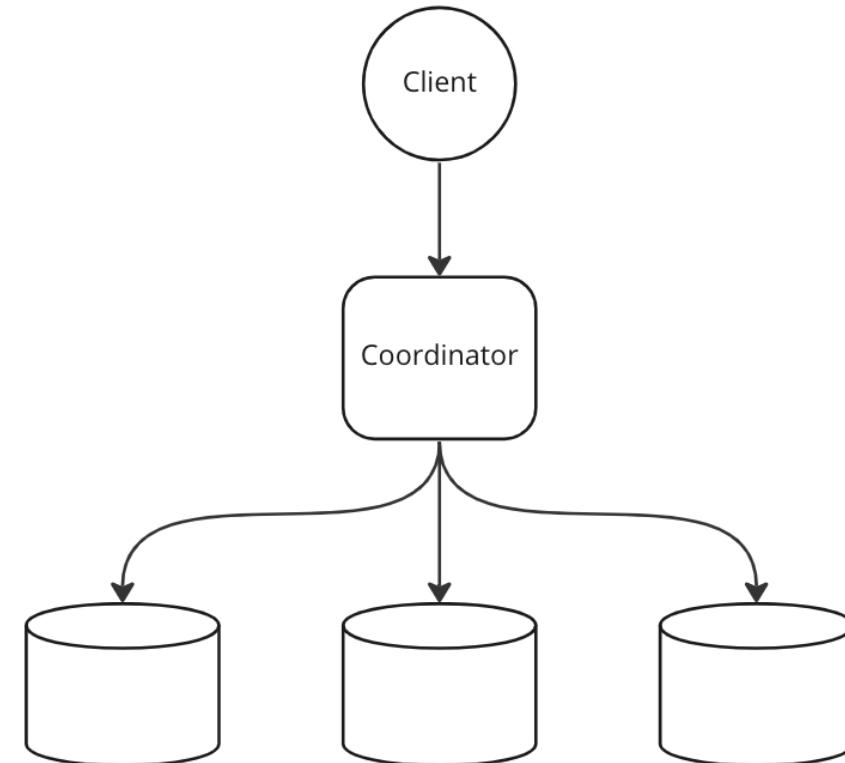
Proxy

- + приложение не знает о шардинге
- дополнительный сетевой узел
- потеря функциональности
- единичная точка отказа



Coordinator

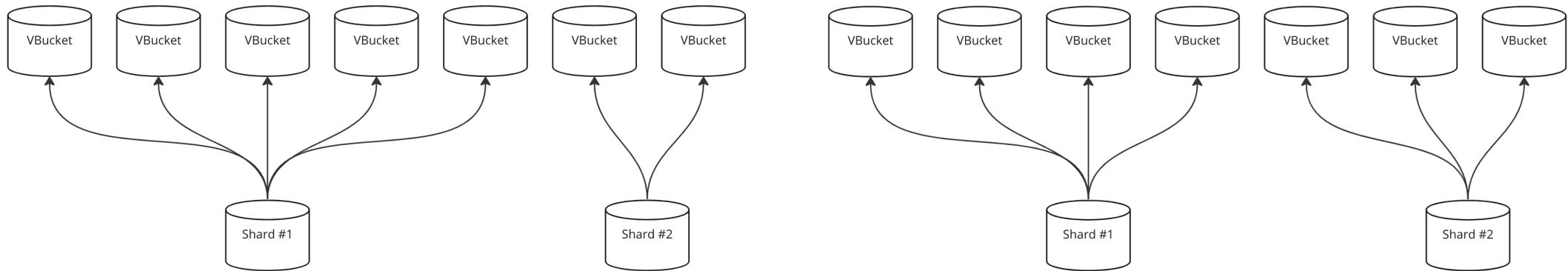
- + кэширование
- + приложение не знает о шардинге
- дополнительный сетевой узел
- инфраструктурная сложность
- единичная точка отказа
- нагрузка



Перебалансировка

Как перенести данные из одного шарда на другой?

Virtual buckets

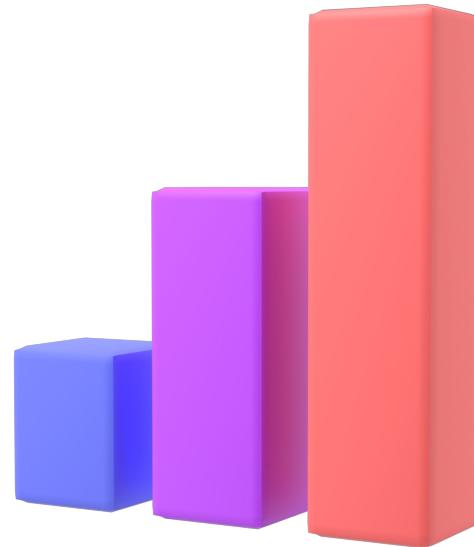


Перебалансировка

1. Только чтение
2. Все данные неизменяемые (*пишем в tgt, читаем из src и tgt*)
3. Логическая репликация с src на tgt, после синхронизации переключаемся на tgt
4. Смешанный подход

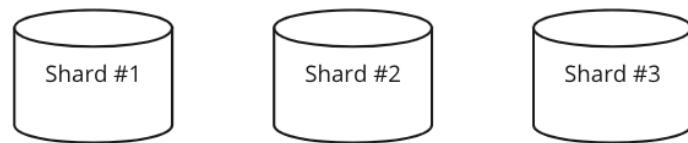
Resharding

- нужно добавить / удалить ноды
- исправление ошибок при выборе стратегии шардирования

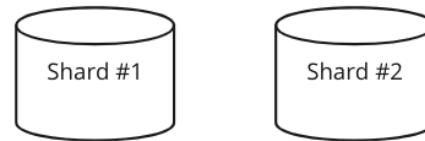


Hashing

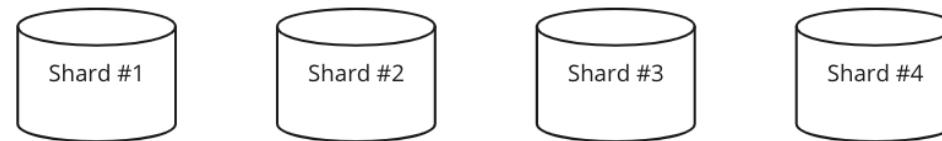
$F(\text{key}) = \text{hash}(\text{key}) \% \text{shards_number}$



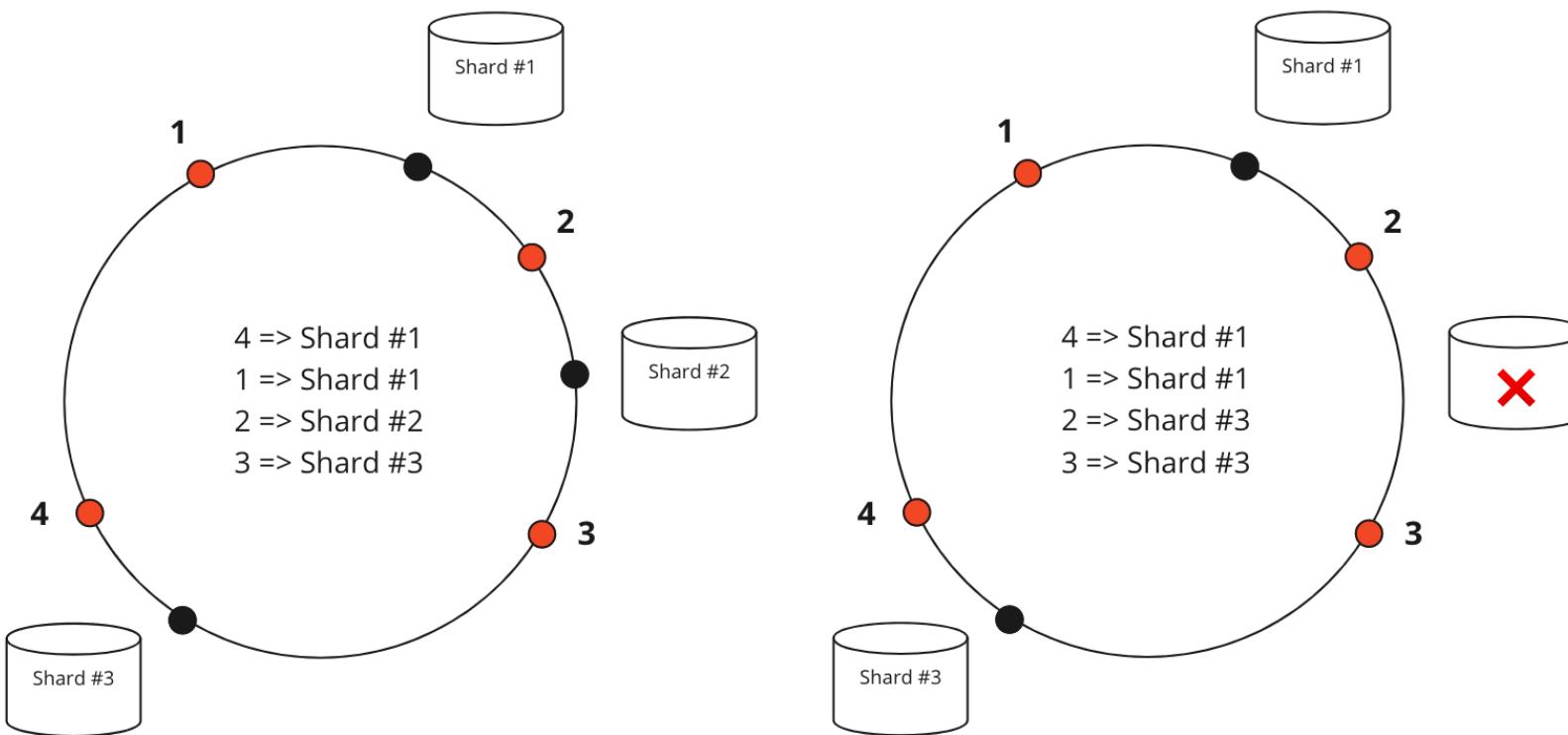
$F(\text{key}) = \text{hash}(\text{key}) \% \text{shards_number}$



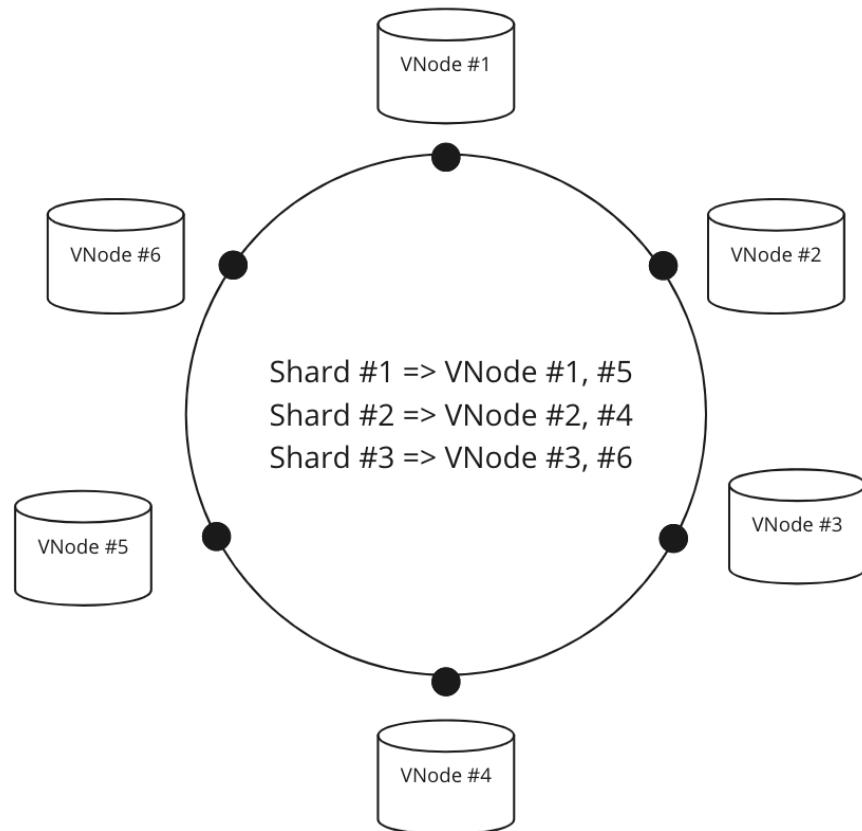
$F(\text{key}) = \text{hash}(\text{key}) \% \text{shards_number}$



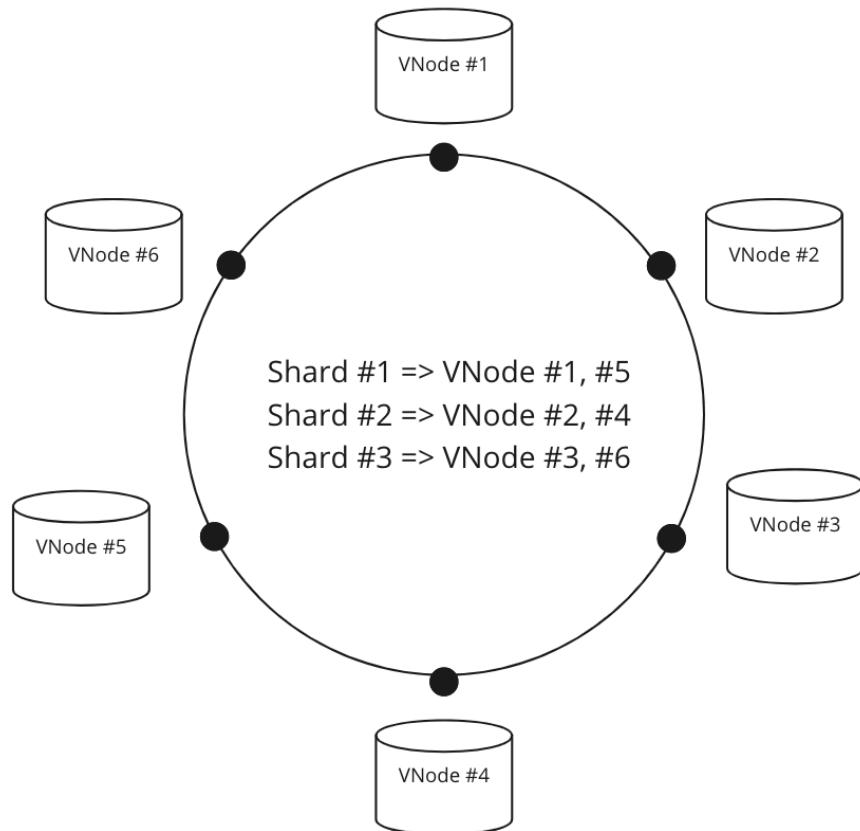
Consistent Hashing



Consistent Hashing

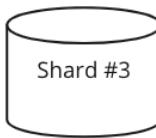
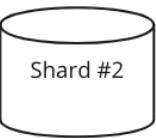
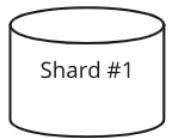


Consistent Hashing



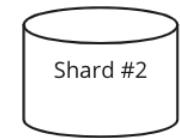
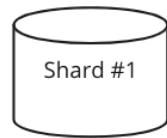
Rendezvous Hashing

$F(123, 1) = \text{hash}(123, 1) = 345$
 $F(123, 2) = \text{hash}(123, 2) = 456$
 $F(123, 3) = \text{hash}(123, 3) = 121$



123

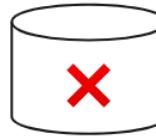
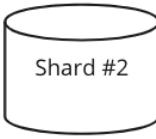
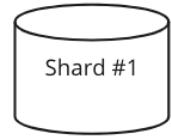
$F(242, 1) = \text{hash}(242, 1) = 233$
 $F(242, 2) = \text{hash}(242, 2) = 124$
 $F(242, 3) = \text{hash}(242, 3) = 434$



123

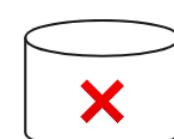
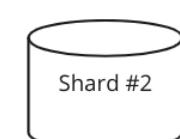
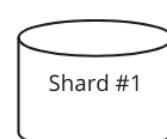
542

$F(123, 1) = \text{hash}(123, 1) = 345$
 $F(123, 2) = \text{hash}(123, 2) = 456$



123

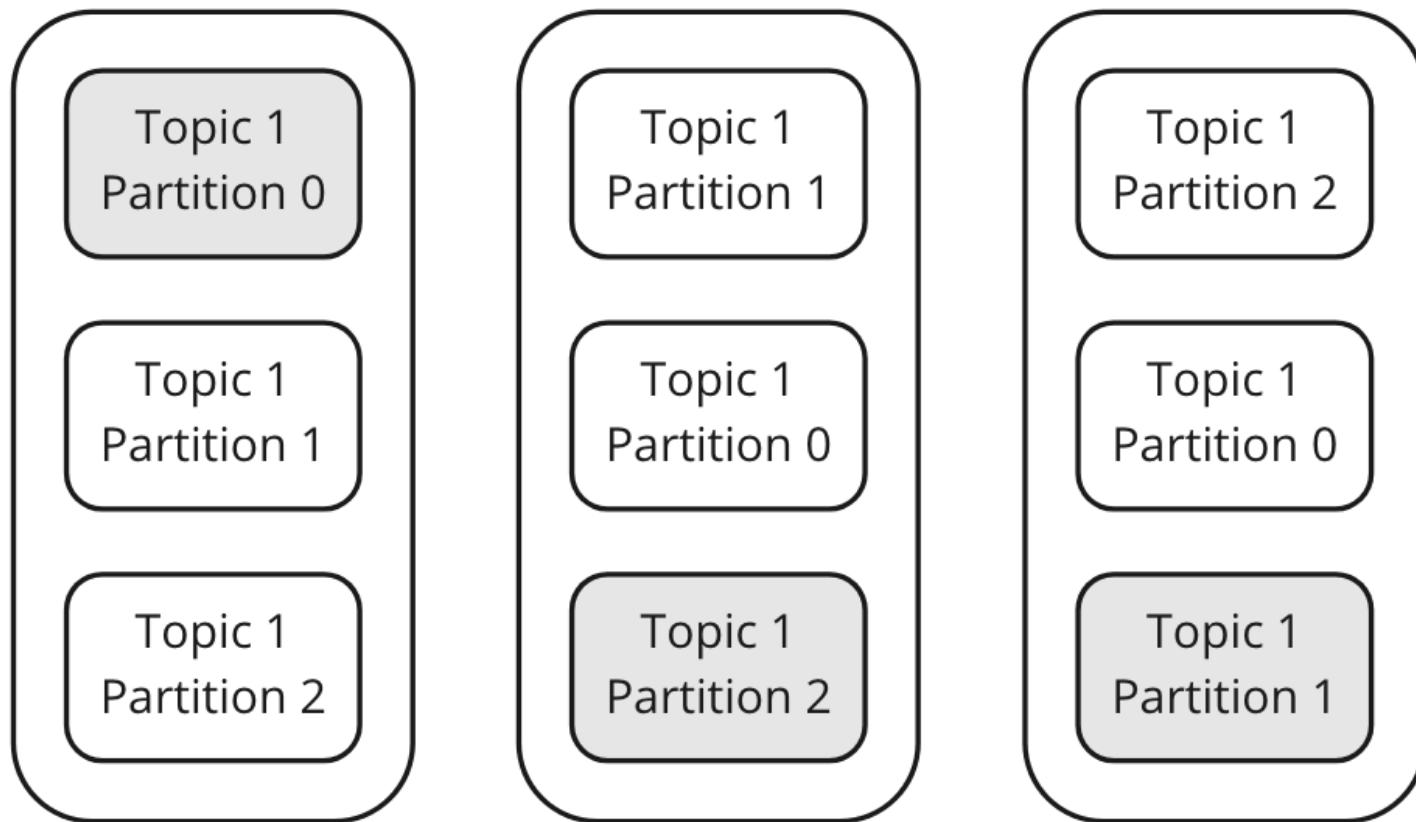
$F(242, 1) = \text{hash}(242, 1) = 233$
 $F(242, 2) = \text{hash}(242, 2) = 124$



542

123

Kafka Cluster



СБЕР БАНК

Логин [вход по номеру телефона](#)

Пароль 

Запомнить меня

Продолжить

Или выберите другой способ входа:

 QR-код

[Зарегистрироваться](#)

[Восстановить доступ](#)

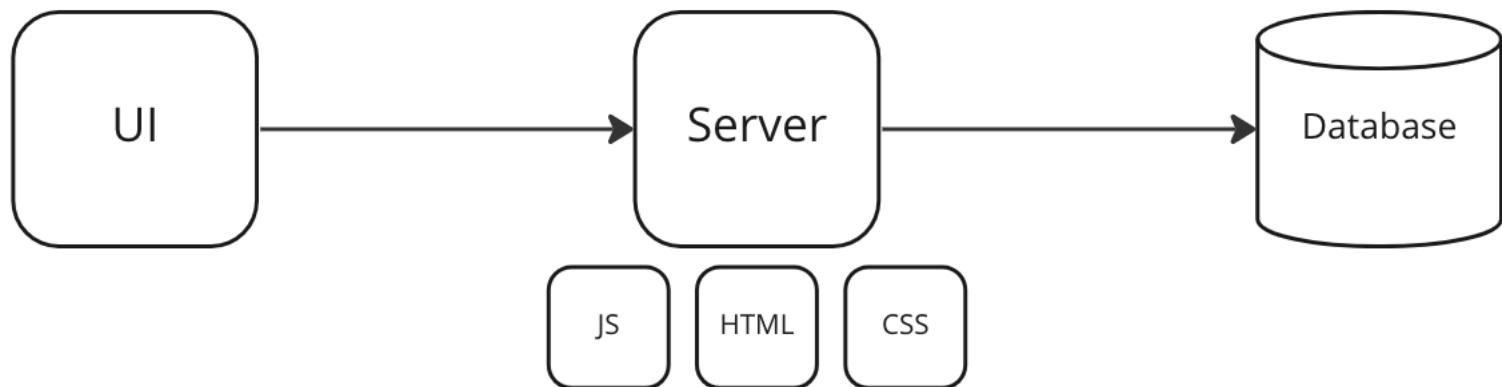
Лучший %

Вклад с доходностью до 9,5% годовых.

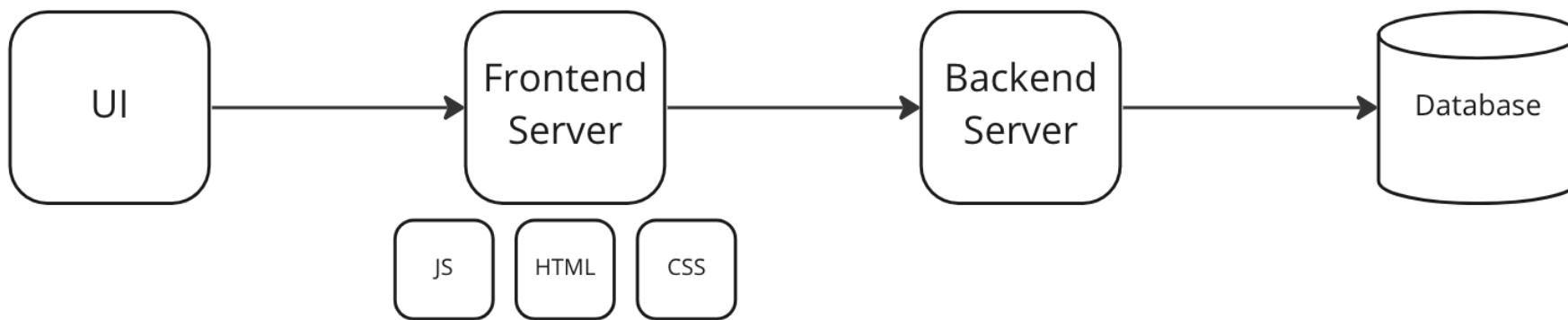
Удобные сроки – от 1 месяца до 3 лет.

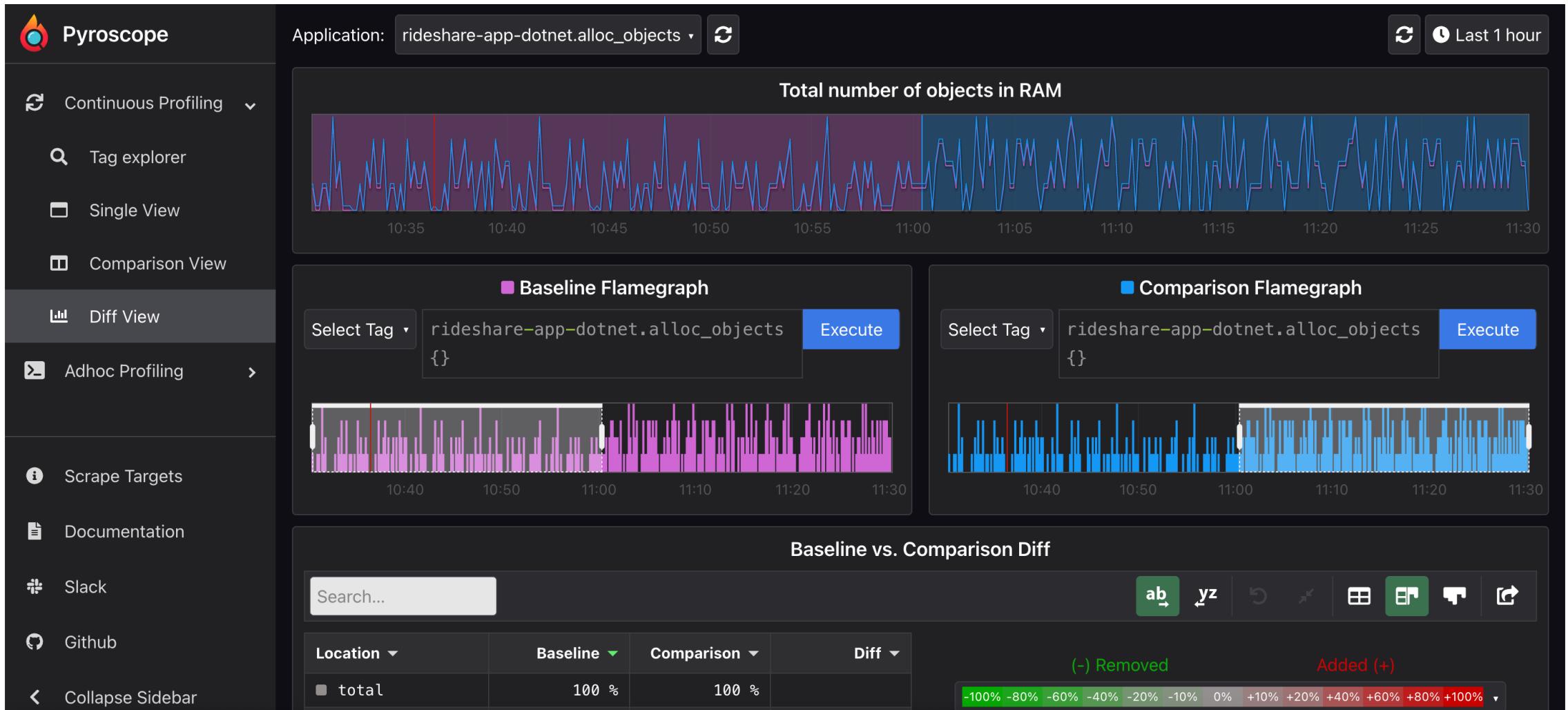


MVP

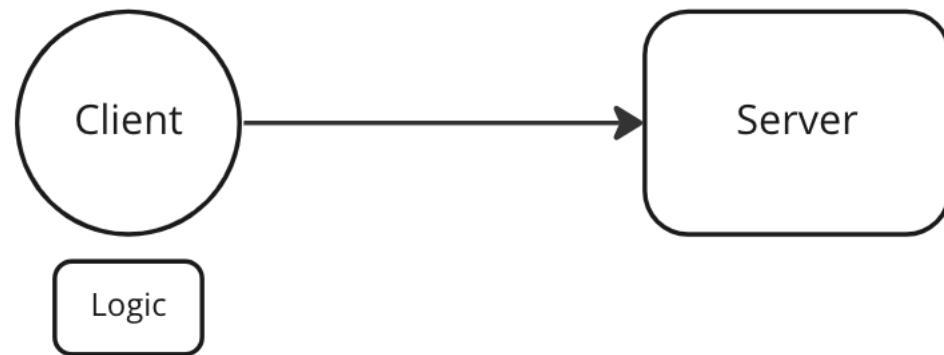


Трехзвенная архитектура





Толстый клиент



VK ВКОНТАКТЕ

Поиск Miyagi — Самурай

Наверх

Блог Разработчикам Для бизнеса Ещё

VK Cloud ✓ Реклама · 18+

VK Kubernetes Conf 2023
21 марта 2023 в 13:00 поговорим о самых больных местах при работе с Kubernetes

VK Cloud VK бизнес

VK Kubernetes Conf '23

Бесплатная регистрация - конференция Kubernetes в России | VK Cloud

mcs.mail.ru Перейти

1 1

Английский юмор сегодня в 9:30

Новости

Фотографии

Видео

Рекомендации

Поиск

Реакции

Обновления

Комментарии

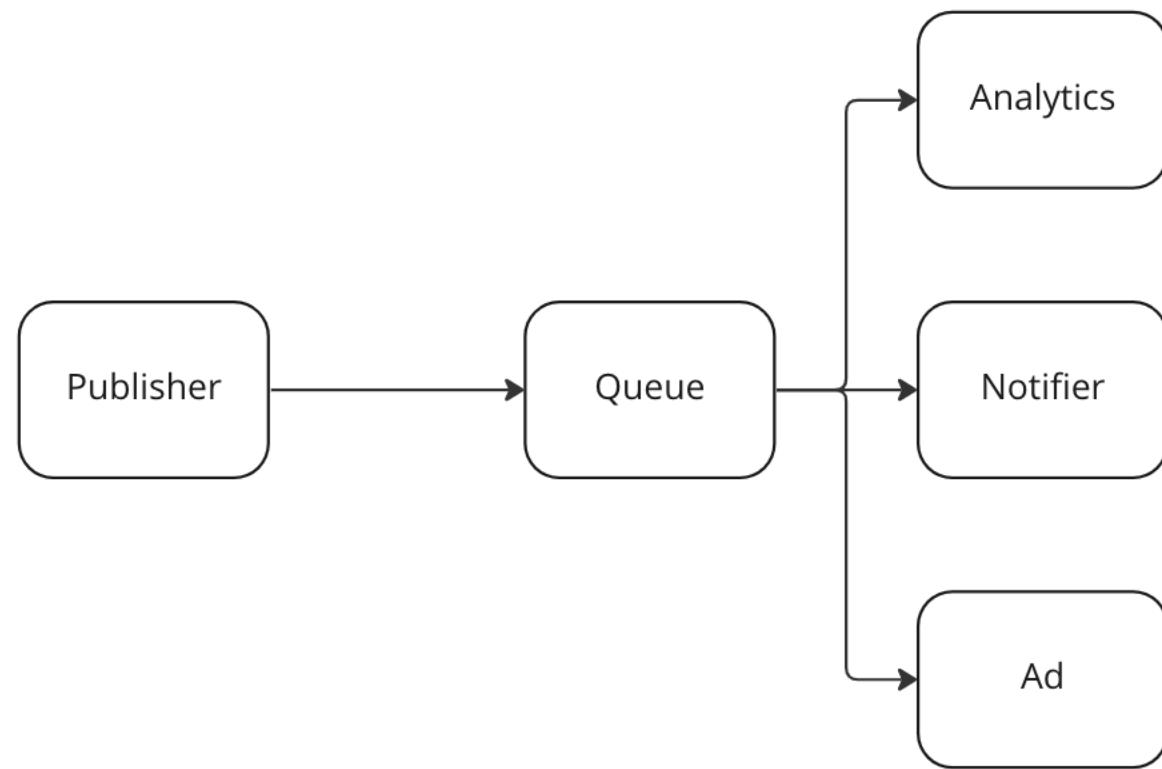
Сначала интересные

Все новости из мира финансов – в одном сервисе А ещё конвертер и курс валют

Открыть

«»

pub / sub



Самое важное про текущую ситуацию: [сводка](#), [телеграм-канал](#)



[Билеты на поезд](#)

[Авиабилеты](#)

[Автобусы](#)

[Отели](#)

[Командировки](#)

[Электрички](#)

[Приключения](#)

[Справочная](#)

[Войти или Зарегистрироваться](#)

Москва — Екатеринбург Пасс. 24 мар, пт

[Изменить поиск](#)

Нижние места
от 4 253 ₽

Тип вагона ▾
Любой

Выберите поезд, чтобы заказать билеты онлайн

Отправление и прибытие по местному времени



22 мар, ср
3 241 ₽

23 мар, чт
3 250 ₽

24 мар, пт
4 253 ₽

25 мар, сб
3 947 ₽

26 мар, вс
4 253 ₽

Россия 002Э

00:35 24 мар, пт

Ярославский вокзал, Москва

11:13 25 мар, сб

Екатеринбург Пасс., Екатеринбург

1д 8ч 38м

ФПК

Маршрут

8.8 2K отзывов

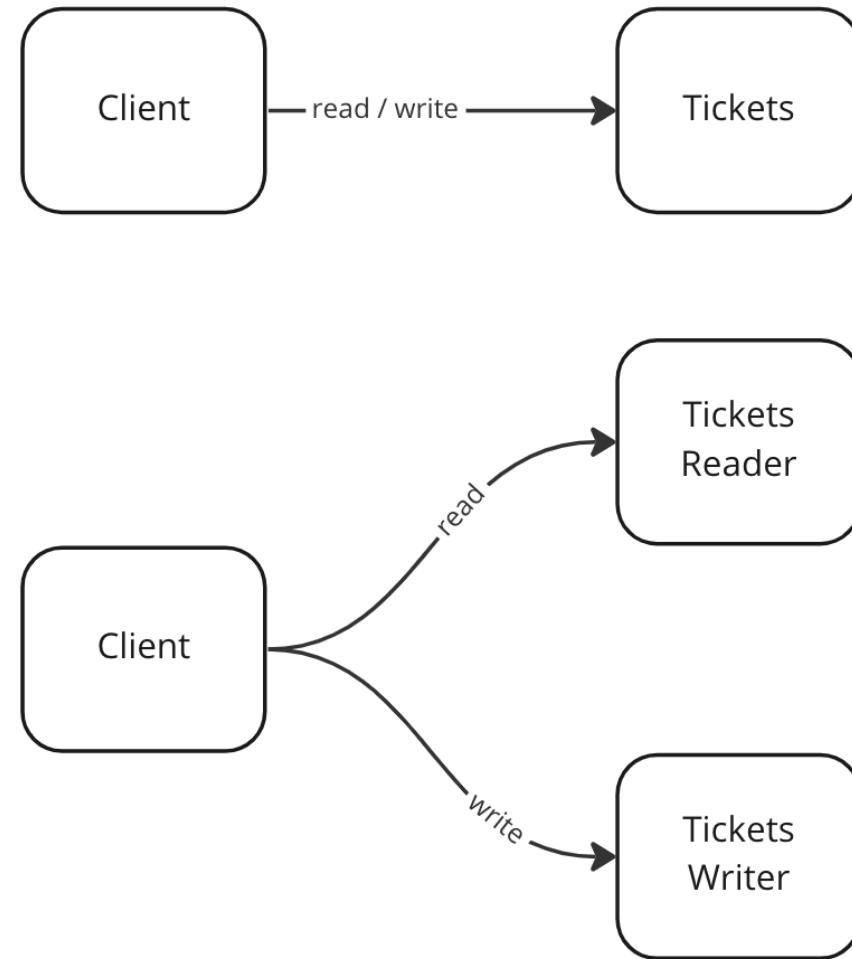


[Выбрать места](#)

Купе 16 ниж, 25 верх **от 4 551 ₽**

Плацкарт 58 ниж, 75 верх **от 4 568 ₽**

CQRS





Авиабилеты

Отели

Для бизнеса ↗



Профиль



Кое-что Ещё



Москва

MOW

Санкт-Петербург

LED

26 апреля, ср



Обратно

1 пассажир
эконом

Найти билеты

Δ Составить сложный маршрут

Длительность пересадок

До 24ч

Если комфорт важнее

Без ночных пересадок



Вылет в Санкт-Петербург

Багаж

Авиакомпании

Альянсы

Время в пути

Аэропорты пересадок

Аэропорты в Москве

Аэропорты в Санкт-Петербурге

Рекомендуемый

2602₽

Багаж +1500₽

Выбрать билет



S7 Airlines

07:30

Москва
26 апр, ср

В пути: 1ч 35м



09:05

Санкт-Петербург
26 апр, ср

LED

Объявление скрыто

Мы используем ваши
ответы, чтобы
подбирать для вас
подходящую рекламу

Самый дешёвый

2097₽

Багаж +1200₽

Выбрать билет



Уральские авиалинии

10:10

Москва
26 апр, ср

В пути: 1ч 35м



11:45

Санкт-Петербург
26 апр, ср

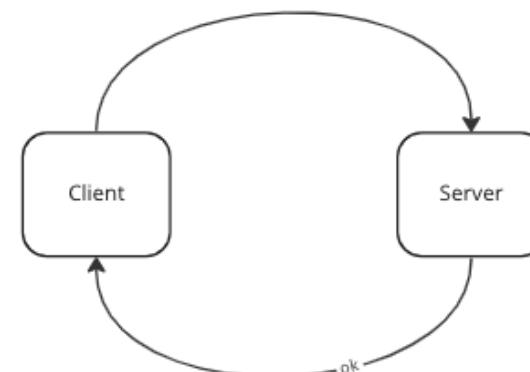
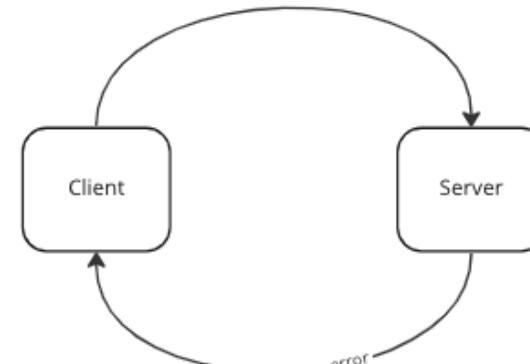
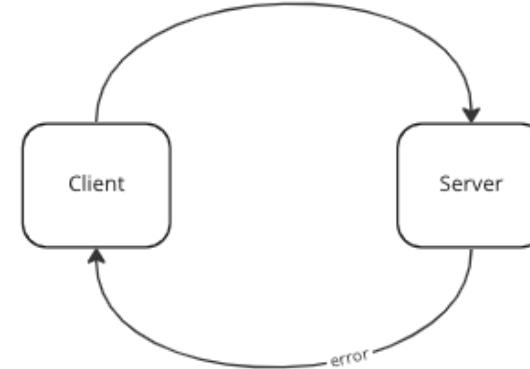
LED

Retries

Нужно найти компромисс между:

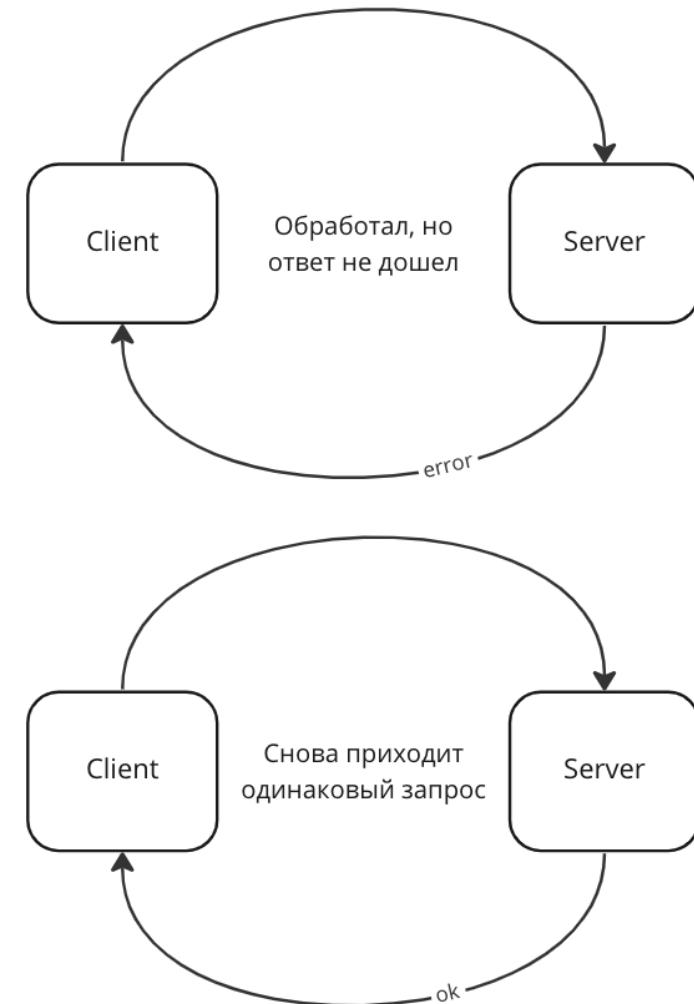
- пользовательский запрос –расшибись, но ответь
- лучше ответить ошибкой, чем перегрузкой сервисов
- при неидемпотентных запросах нельзя повторять

определенные типы запросов



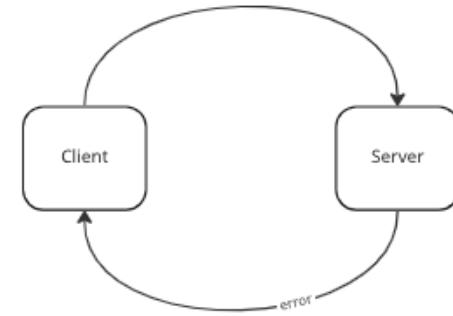
Идемпотентность

Свойство объекта или операции (GET, PUT, DELETE)
при повторном применении операции к объекту
давать тот же результат, что и при первом

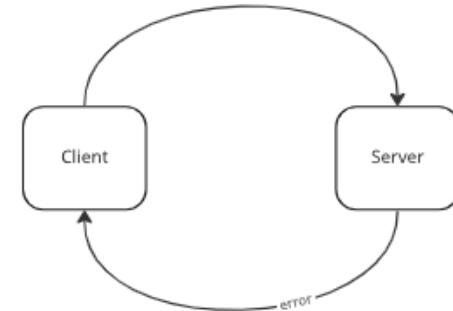


backoff

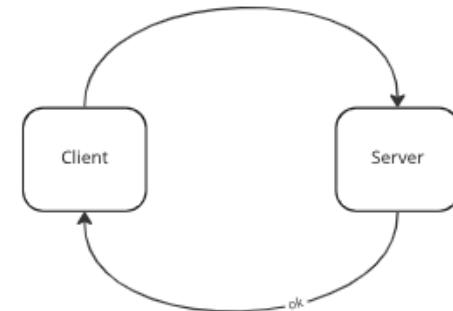
Повтор через фиксированный интервал
времени и экспоненциальное откладывание



Подождать 100мс

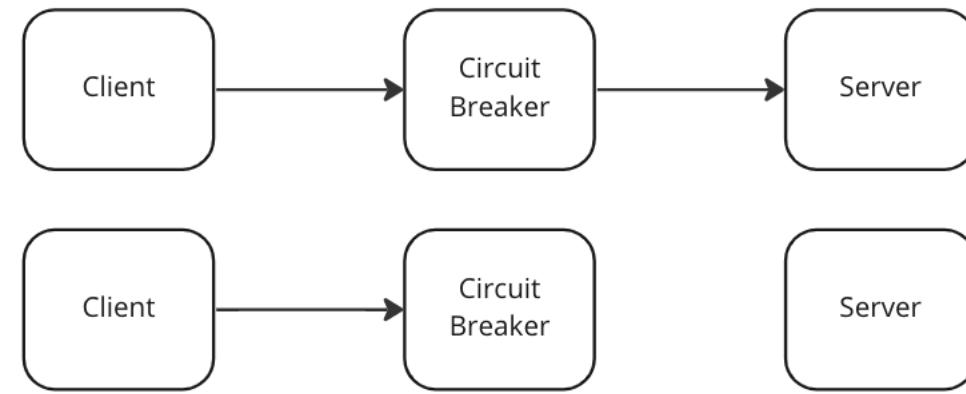


Подождать 200мс



Circuit Breaker

1. Closed
2. Open
3. Half - open



Studio

Поиск на канале

создать

Загрузка видео

Ваш канал
Владимир Балун

Главная

Контент

Аналитика

Комментарии

Субтитры

Авторские права

Настройки

Отправить отзыв

Коммент... % "Нравится"

Загрузка видео

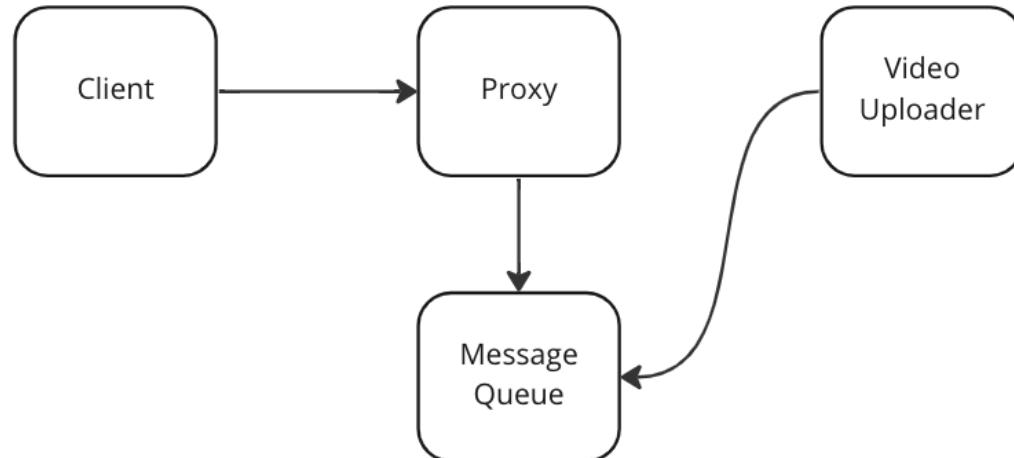
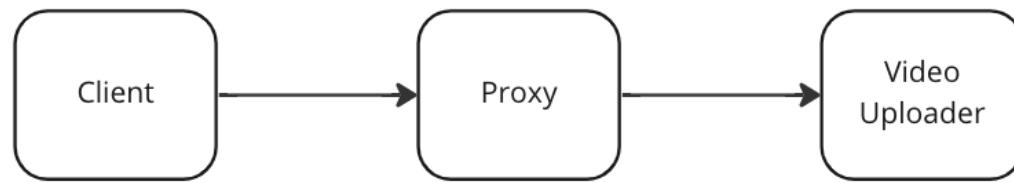
Перетащите файлы сюда или нажмите кнопку ниже, чтобы выбрать их на компьютере.

Пока вы не опубликуете видео, доступ к ним будет ограничен.

ВЫБРАТЬ ФАЙЛЫ

Добавляя видео, вы принимаете [Условия использования](#) и [правила сообщества YouTube](#).
Также вы обязуетесь соблюдать авторские права и конфиденциальность данных других пользователей. [Подробнее...](#)

Отложенные задачи





Банк Бизнес Касса Инвестиции Сим-карта Страхование Путешествия Город Долями

Личный кабинет



Портфель Каталог Пульс Аналитика Академия Скринер Новое Терминал Настройки 1

Все продукты



Yandex YNDX

Доходность за полгода Сектор

-7,71%

Телекоммуникации



Цена акции

1 987 ₽

Продать

Купить

Обзор Пульс Прогнозы Показатели Новости Идеи События

★ В избранном

M5 M15 M30 Ч1 Ч4 Д Н Мес



2150,0

2100,0

2050,0

2000,0

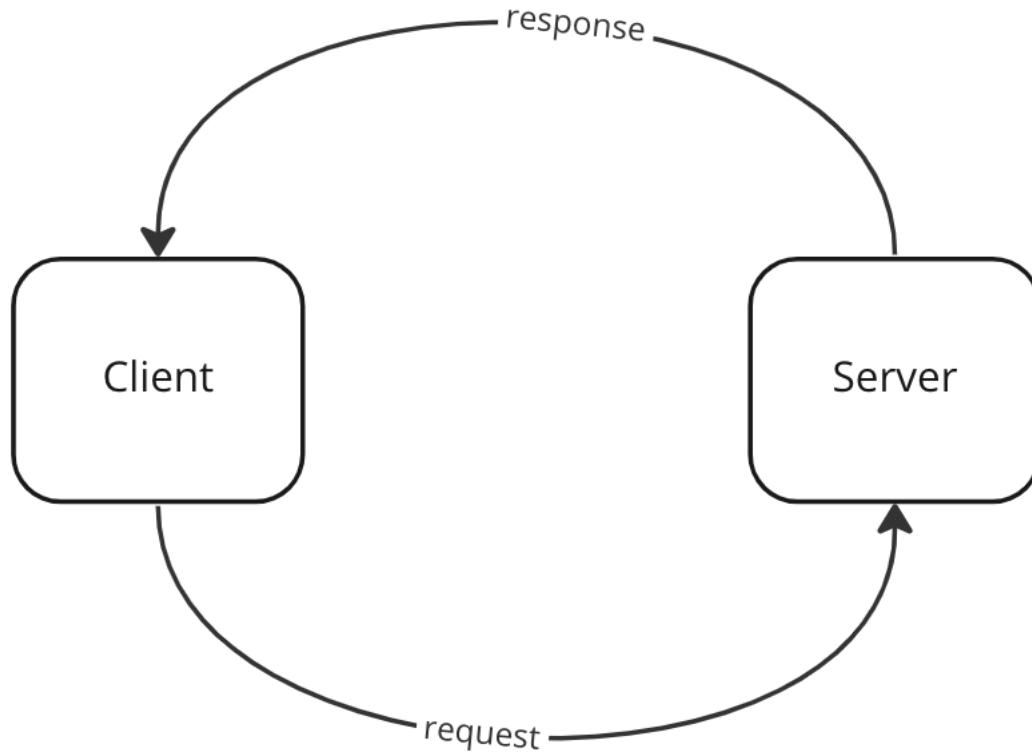
1950,0

Google Chrome

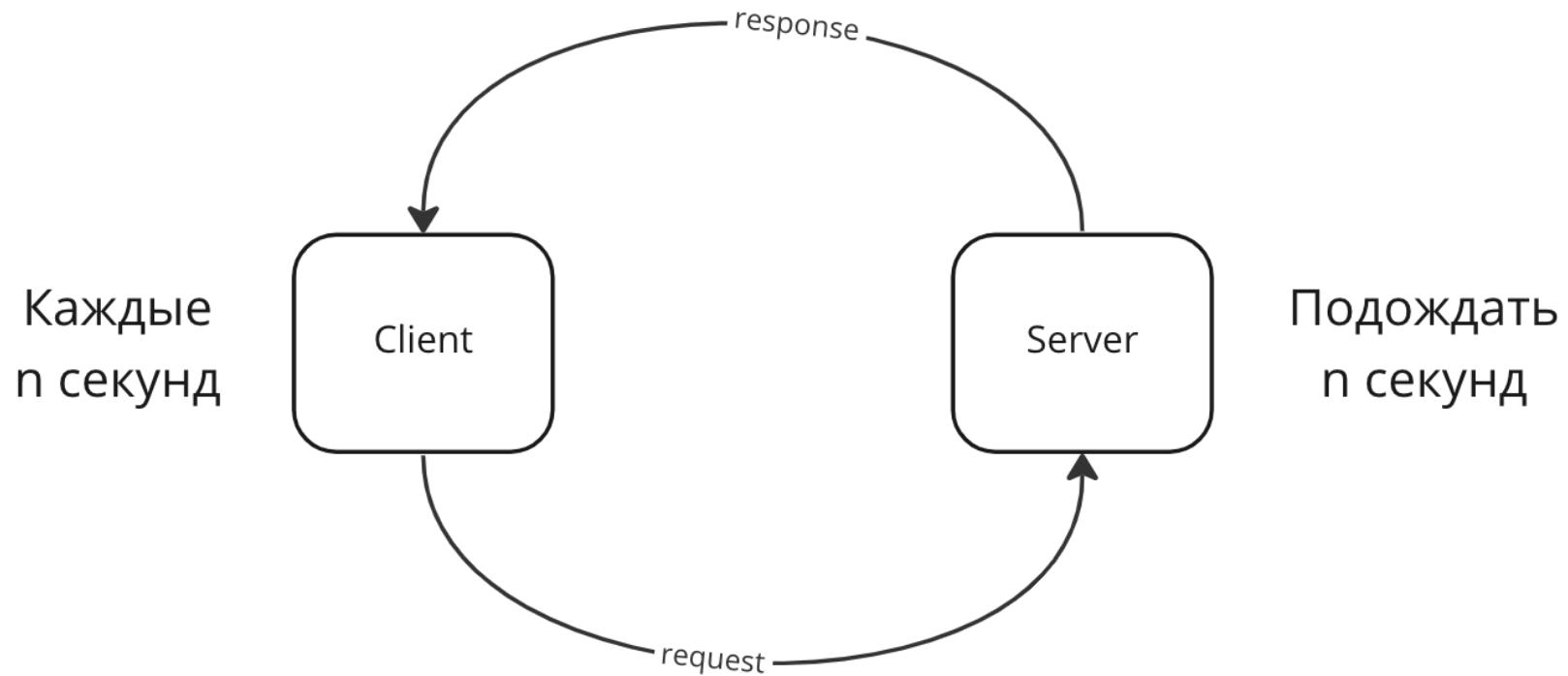


Polling

Каждые
n секунд



Long Polling



Streaming



Яндекс Go Такси

Пользователям ▾ Водителям ▾ Бизнесу ▾ Партнёрам ▾

vladimirbalun

Откуда ЦСКА

Куда? • 26 мин Красная площадь, 1

+ Добавить остановку

Эконом 467₽ Комфорт 521₽ Комфорт+ 645₽

Business 995₽ Детский 541₽ Минивэн 648₽

Способ оплаты MasterCard 9... >

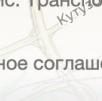
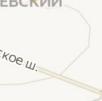
Промокод >

Заказать высокий спрос

+7 (495) 999 99 99 > Заказ такси по телефону



7 МИН



© 2011-2023 ООО «Яндекс.Такси». Яндекс Go — информационный сервис. Транспортные и иные услуги оказываются партнерами сервиса.

Тарифы Партнёры Пользовательское соглашение Лицензионное соглашение

Россия (Русский) Поддержка ?

Graceful Degradation

Payment

Navigator

YPlus

Advertisement

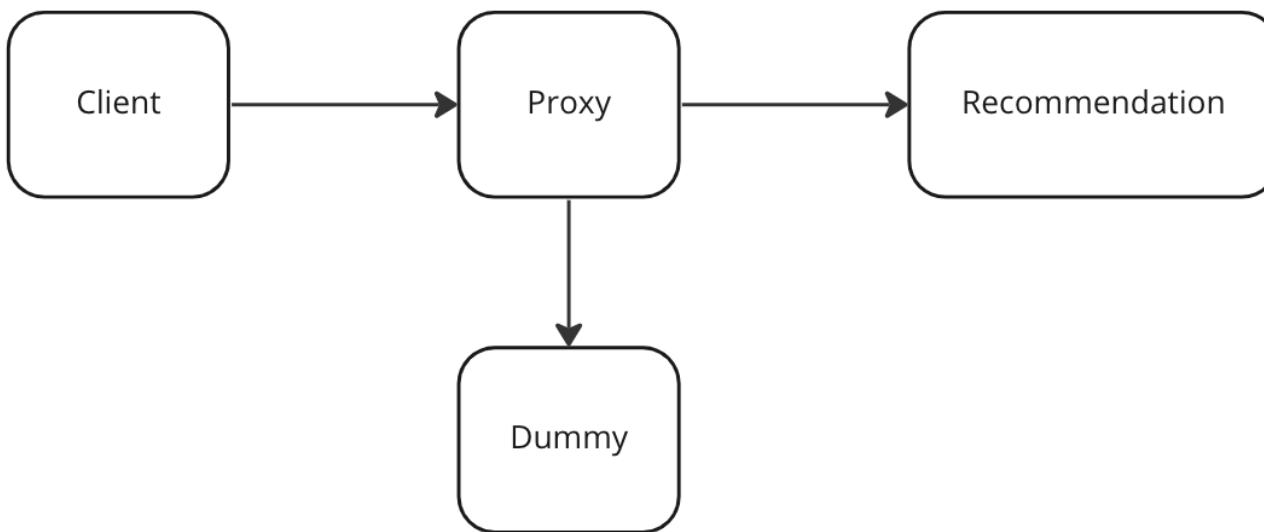
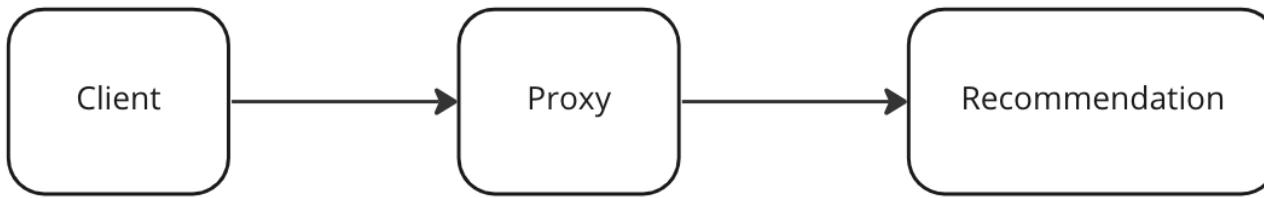
Driver

Customer

Calculation

Load

fallback

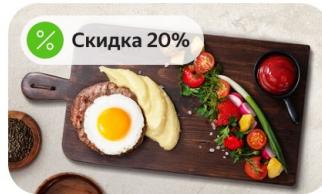




Offers



Tanuki
★ 4.8 Great ₽₽₽ 5%



Supergood
★ 4.8 Great ₽₽₽

25% for first order



Zotman Pizza
★ 4.9 Great ₽₽₽ 5%



АндерСон
★ 4.8 Great ₽₽₽

15% for first order



Самый самолёт
★ 4.7 Good ₽₽₽



Highlighted



Aroma
PROMO ★ 4.8 Great ₽₽₽



Грузины здесь
PROMO ★ 4.9 Great ₽₽₽



Supergood
PROMO ★ 4.8 Great ₽₽₽

25% for first order



Zotman Pizza
PROMO ★ 4.9 Great ₽₽₽

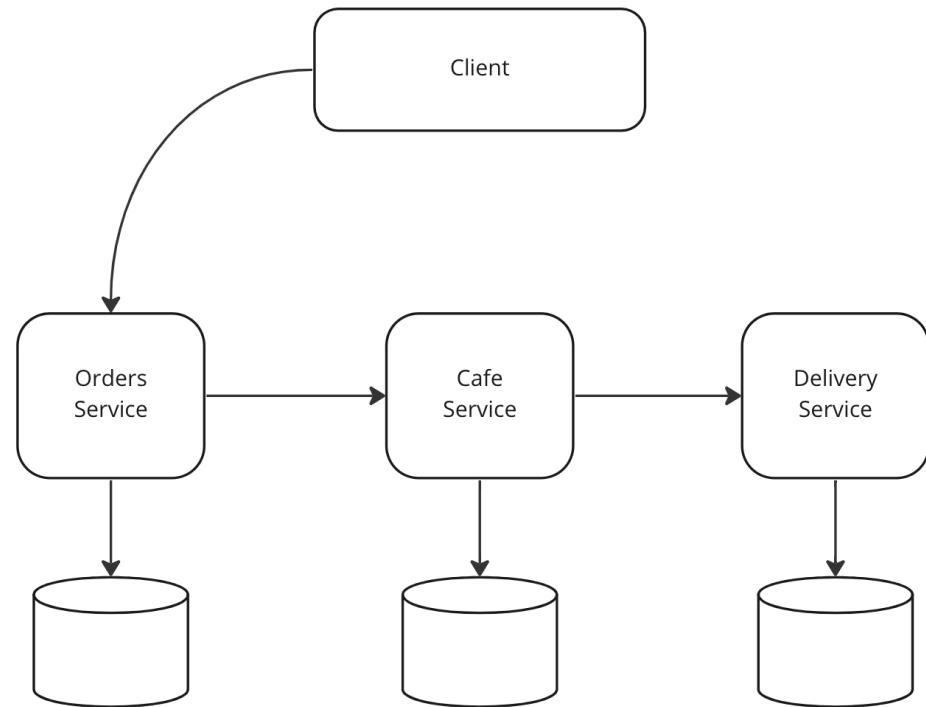


Izumi
PROMO ★ 4.8 Great ₽₽₽



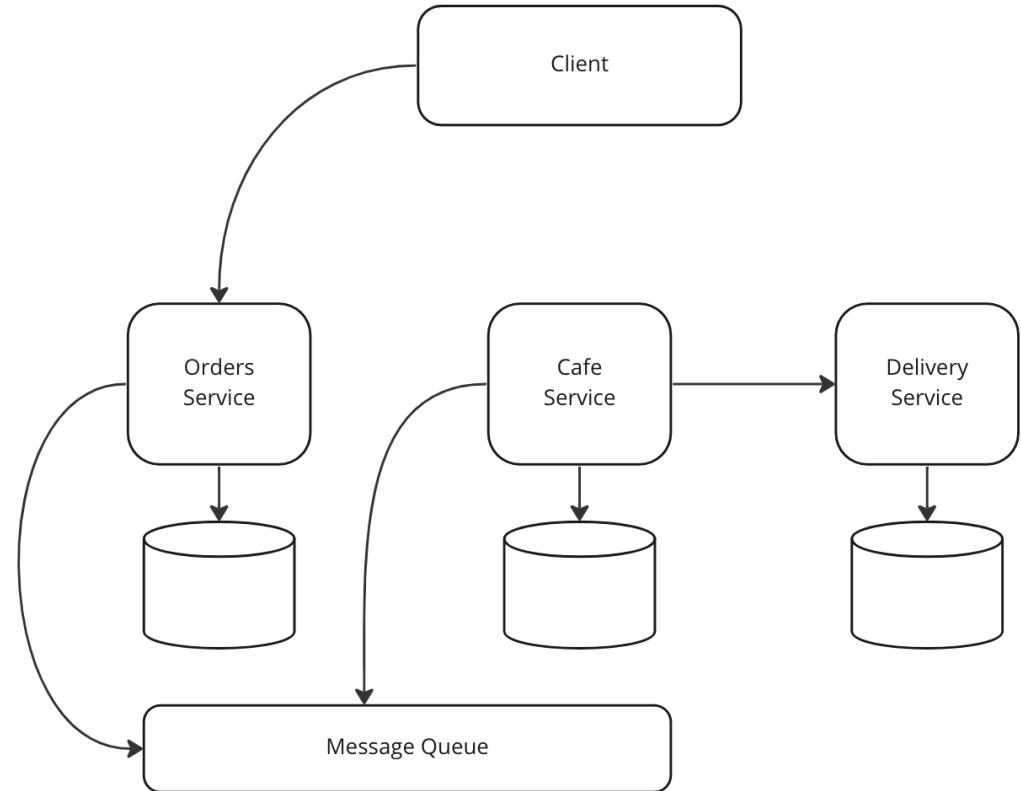
EDA

Event – driven architecture



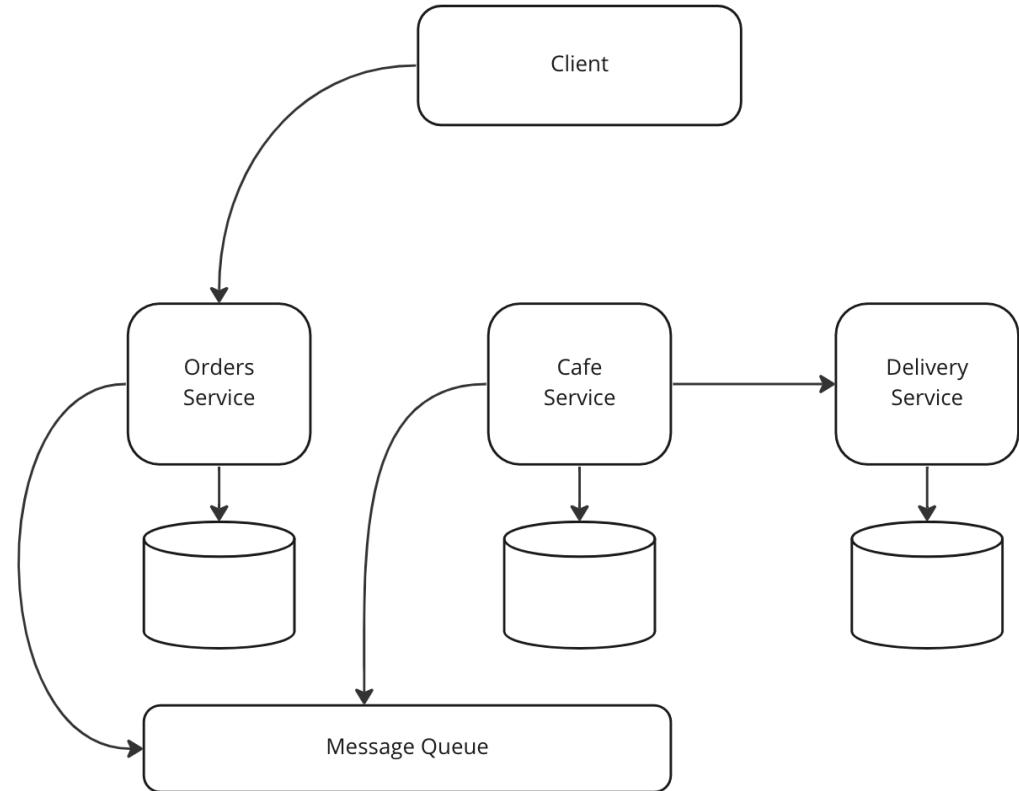
Event Notification

1. Обрабатываем запрос
2. Отпускаем клиента
3. Используем события



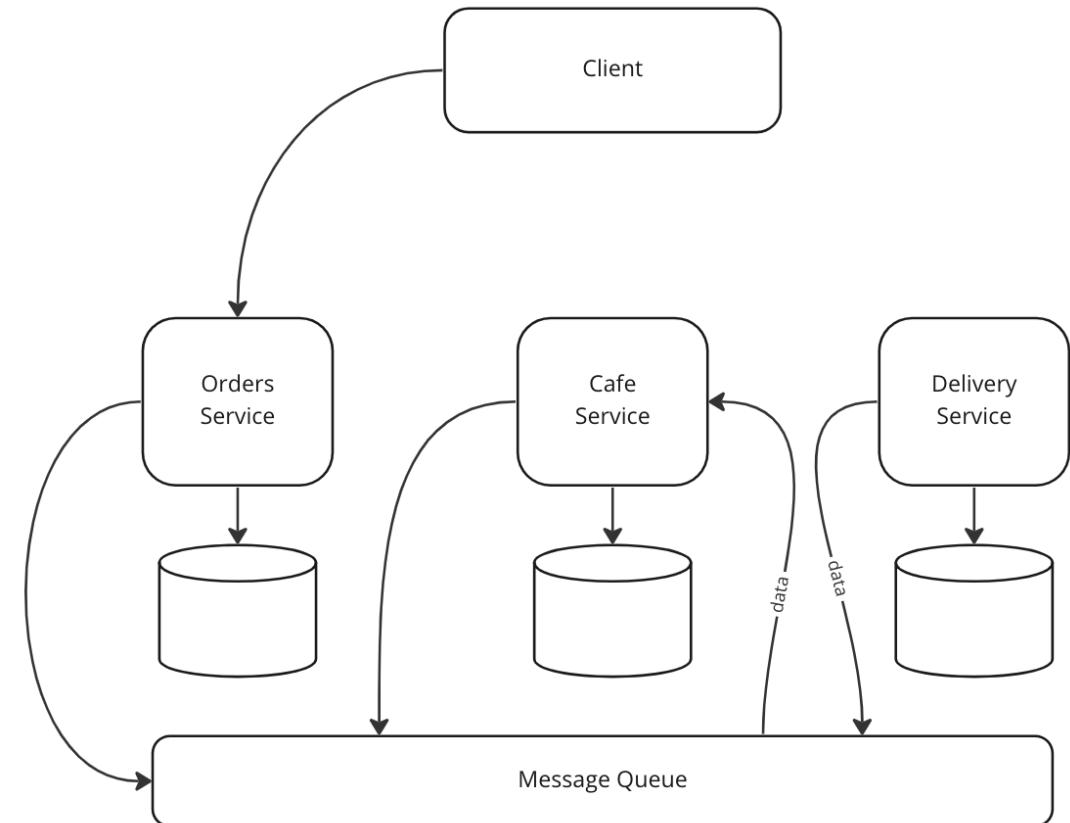
Event Notification

1. Обрабатываем запрос
2. Отpusкаем клиента
3. Используем события



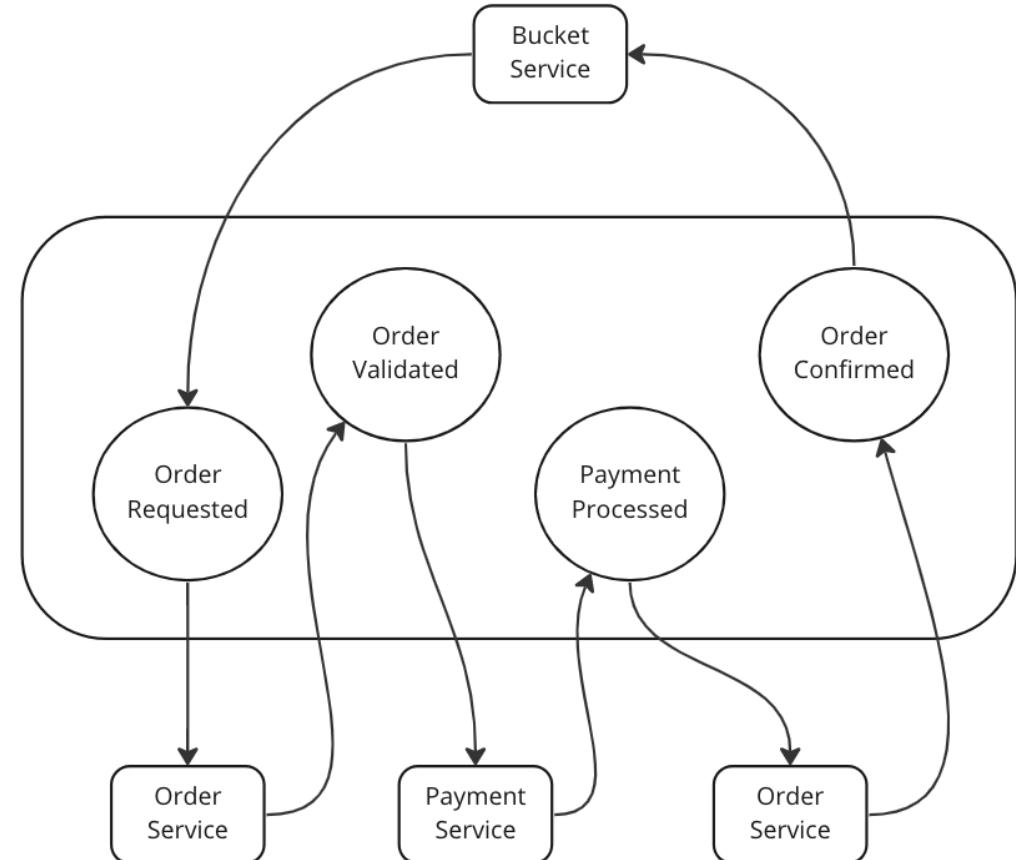
State Transfer

1. Обрабатываем запрос
2. Отпускаем клиента
3. Используем события
4. Реплицируем данные



Event Collaboration

Используется шина для
коммуникации между сервисами





Поиск Яндекса

Найти

Владимир Балун
vladimirbalun@mail.ru

Входящие 8

Написать письмо

Создать почту



ВКонтакте

Вы могли пропустить



Сообщения

1



Облако

Место для ваших файлов

Надежная защита

Облако – это персональное хранилище, где ваши фото и файлы находятся в безопасности

Перейти

реклама

билайн Меняй условия Тарифа UP

Скидки для семьи

Узнать больше

Новости Спецоперация Москва Спорт Леди Авто Кино Hi-Tech Игры Дети Здоровье ...

[Лавров: Россия больше не позволит Западу взрывать газопроводы](#)

Россия впредь будет ориентирована на надежных партнеров, в том числе Китай

[Песков ответил на предложение о военном положении в регионах РФ](#)[В Москве началась церемония прощания с Глебом Павловским](#)[Глава села рассказал, как диверсанты атаковали Брянскую область](#)[Эксперты спрогнозировали доллар выше 200 рублей в 2025 году](#)[Пригожин заявил о практически полном окружении Артемовска](#)[Россия уничтожила базу диверсантов-подводников ВСУ в Николаеве](#)[В Совфеде предложили проект по «переселению» заключенных](#)

Москва



+1°

Снег

Влажность 72%

Вечером

Ночью

Утром



-6°



-12°



-7°

\$ 75,44

+0,12

€ 80

+0,55

₴ 85,82

-0,17

Тегирование кэша

ID	Data
12312	News #1 <small>weather</small>
6423	News #2
1312	News #3 <small>currency, weather</small>

Tags

currency
weather

ID	Data
moscow	<small>weather</small> data



See what's in it for you.

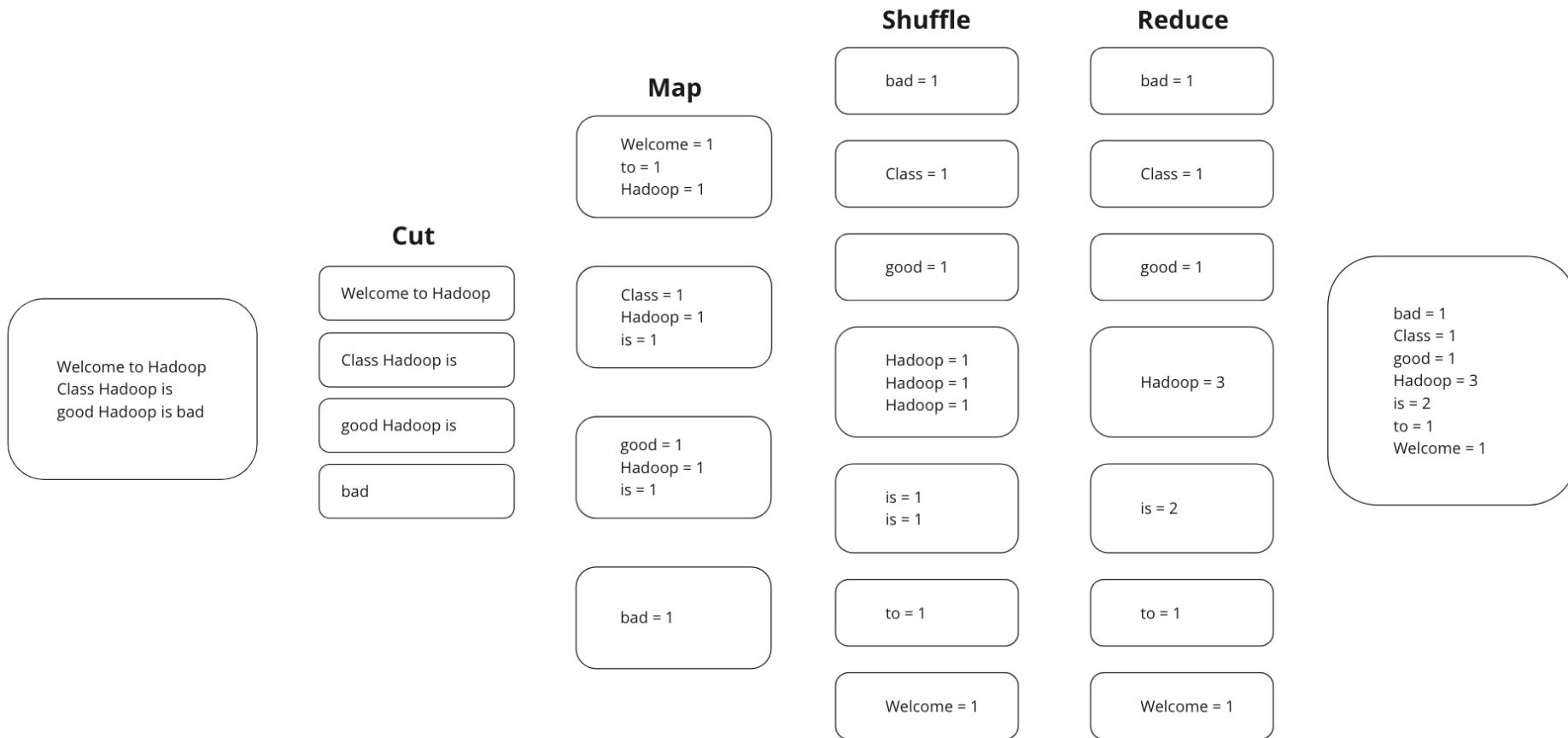


Unlock customer-centric measurement.

Understand how your customers interact across your sites and apps, throughout their entire lifecycle.



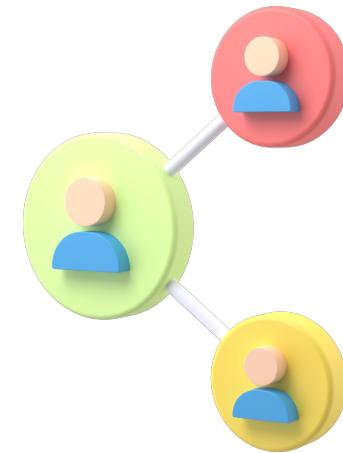
MapReduce



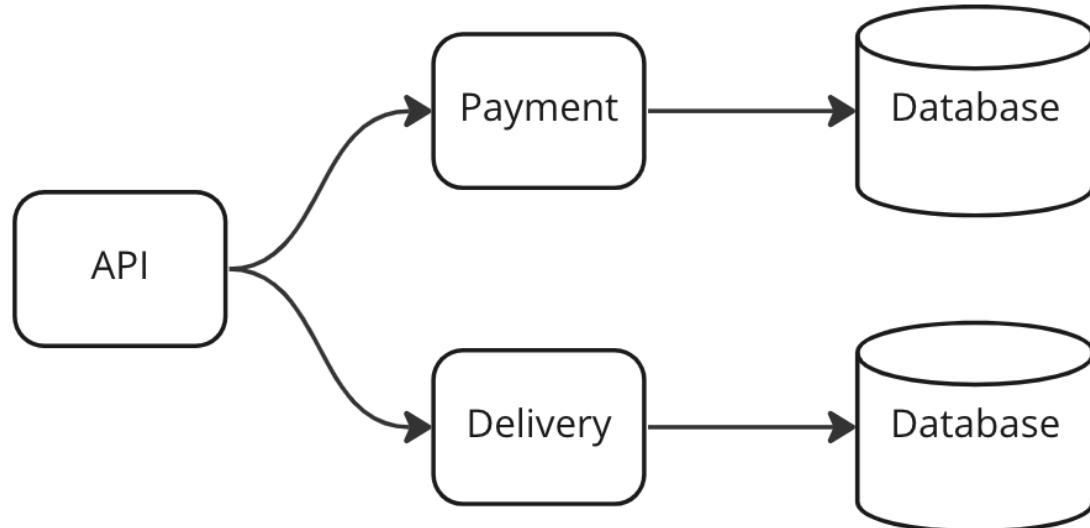
Консенсус

Необходимость нескольких узлов согласовать

Определенные действия между собой

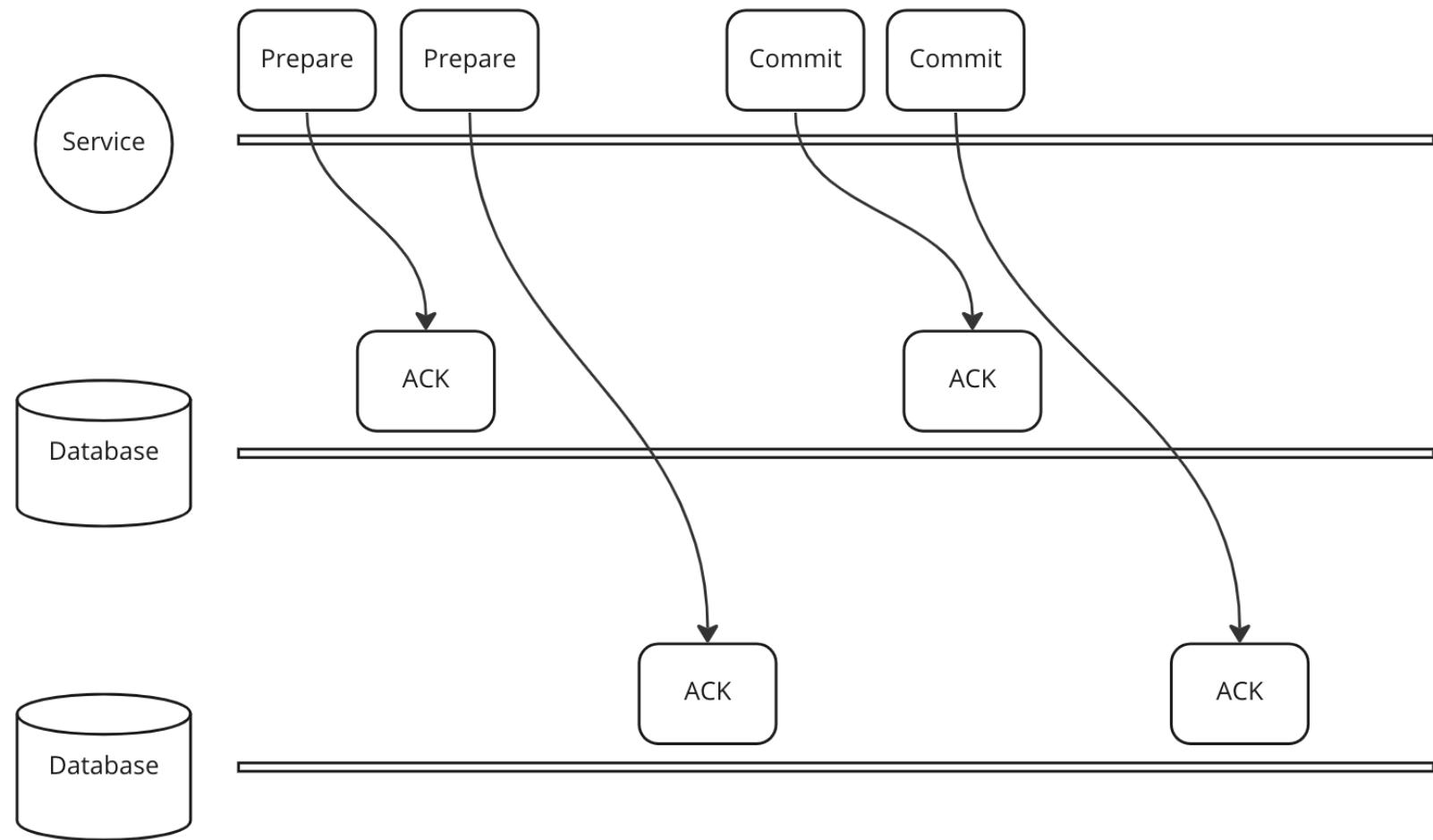


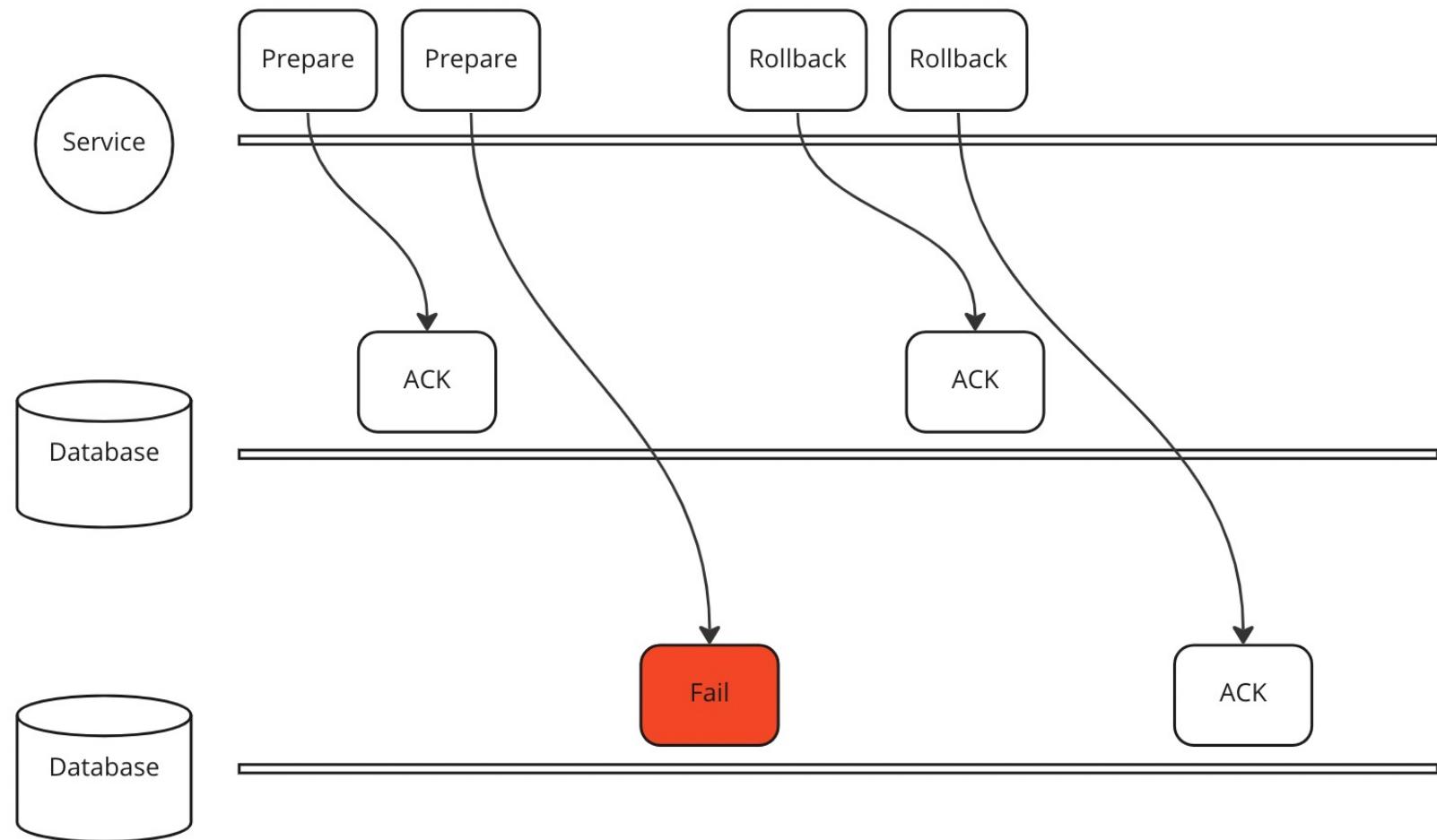
Распределенные транзакции

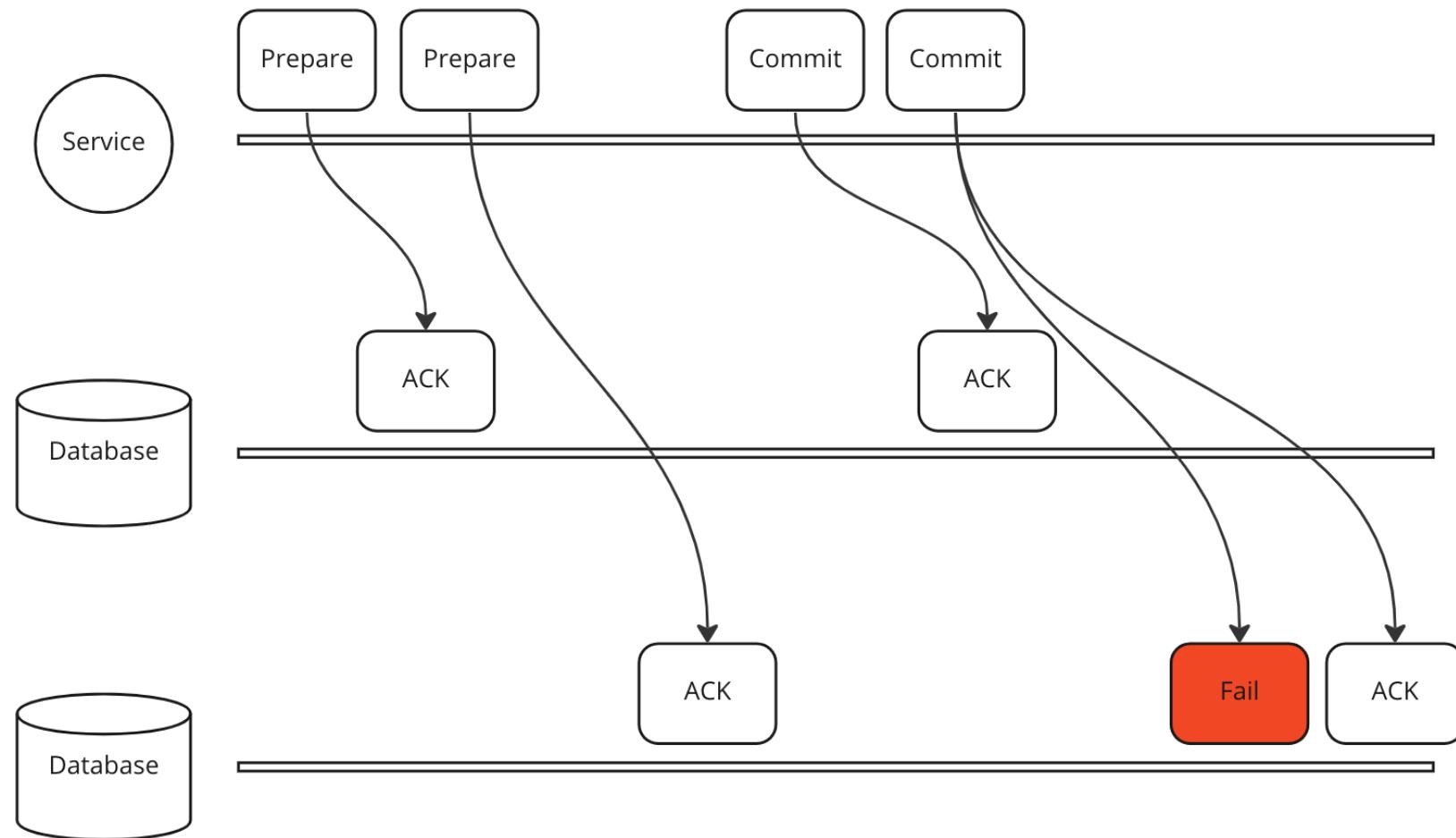


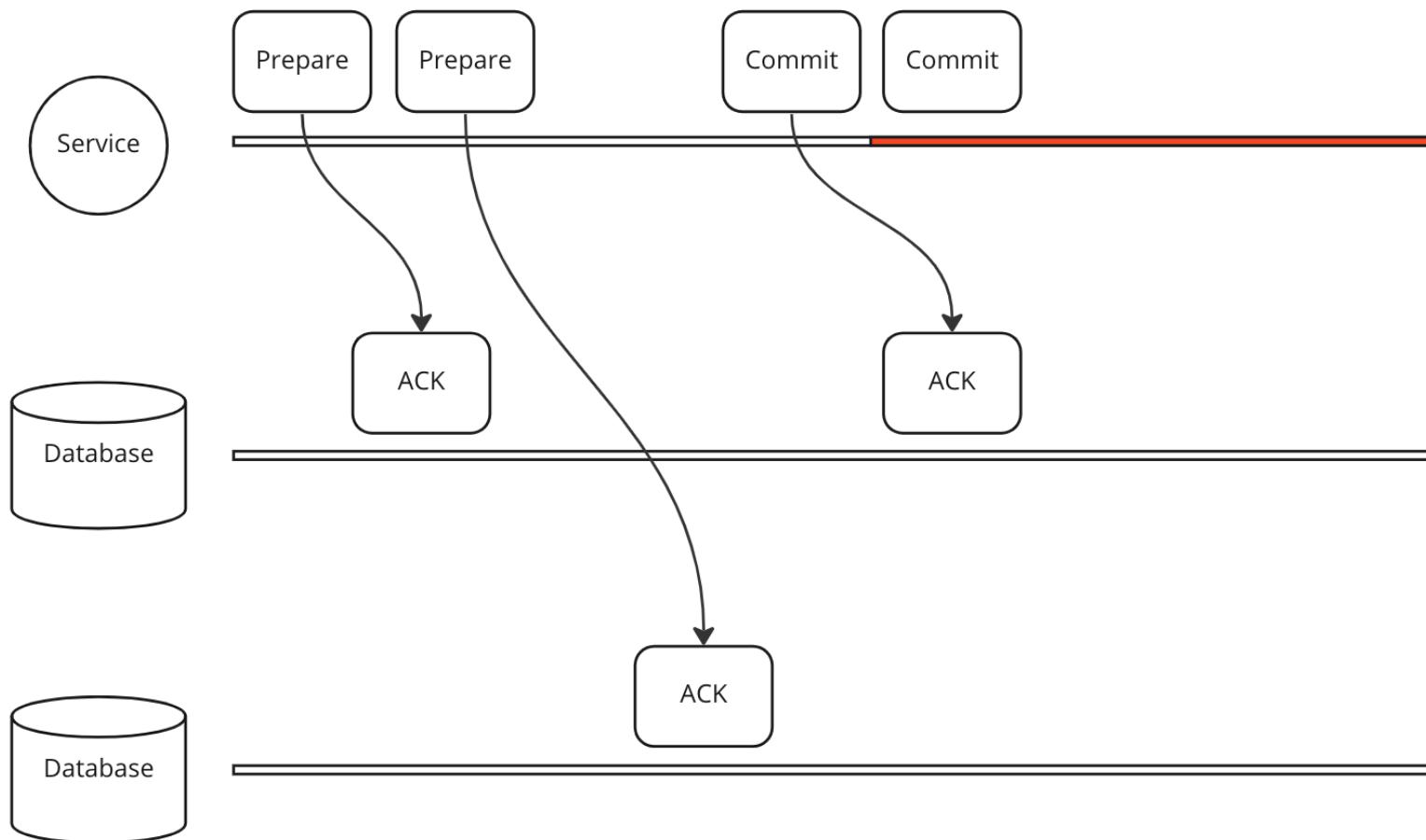
Двухфазная фиксация (2РС)

В ЗАГСе перед тем, как обручить супругов у них спрашивают согласия и только затем обручают (итого получается две фазы – подготовка и фиксация)





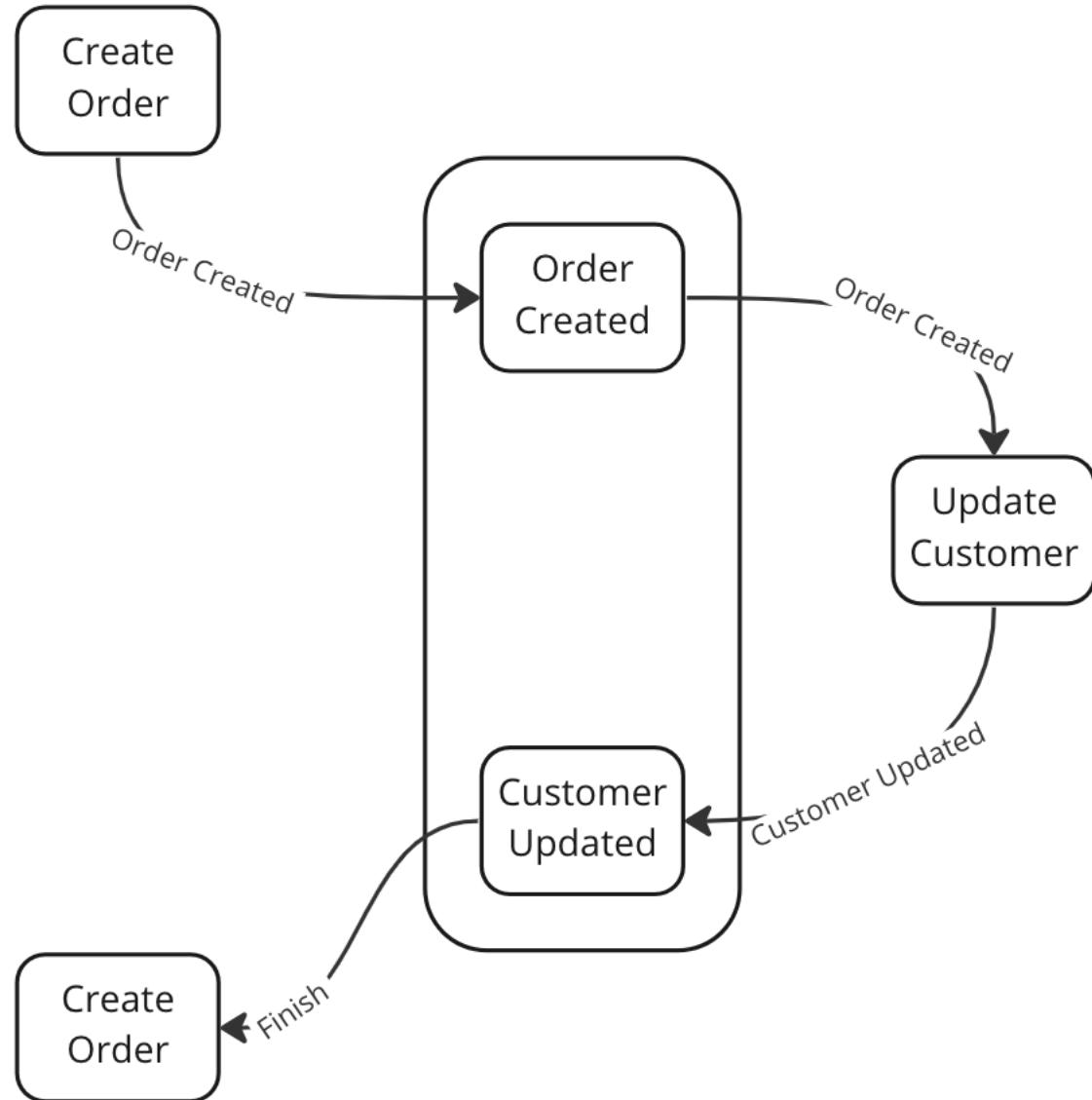


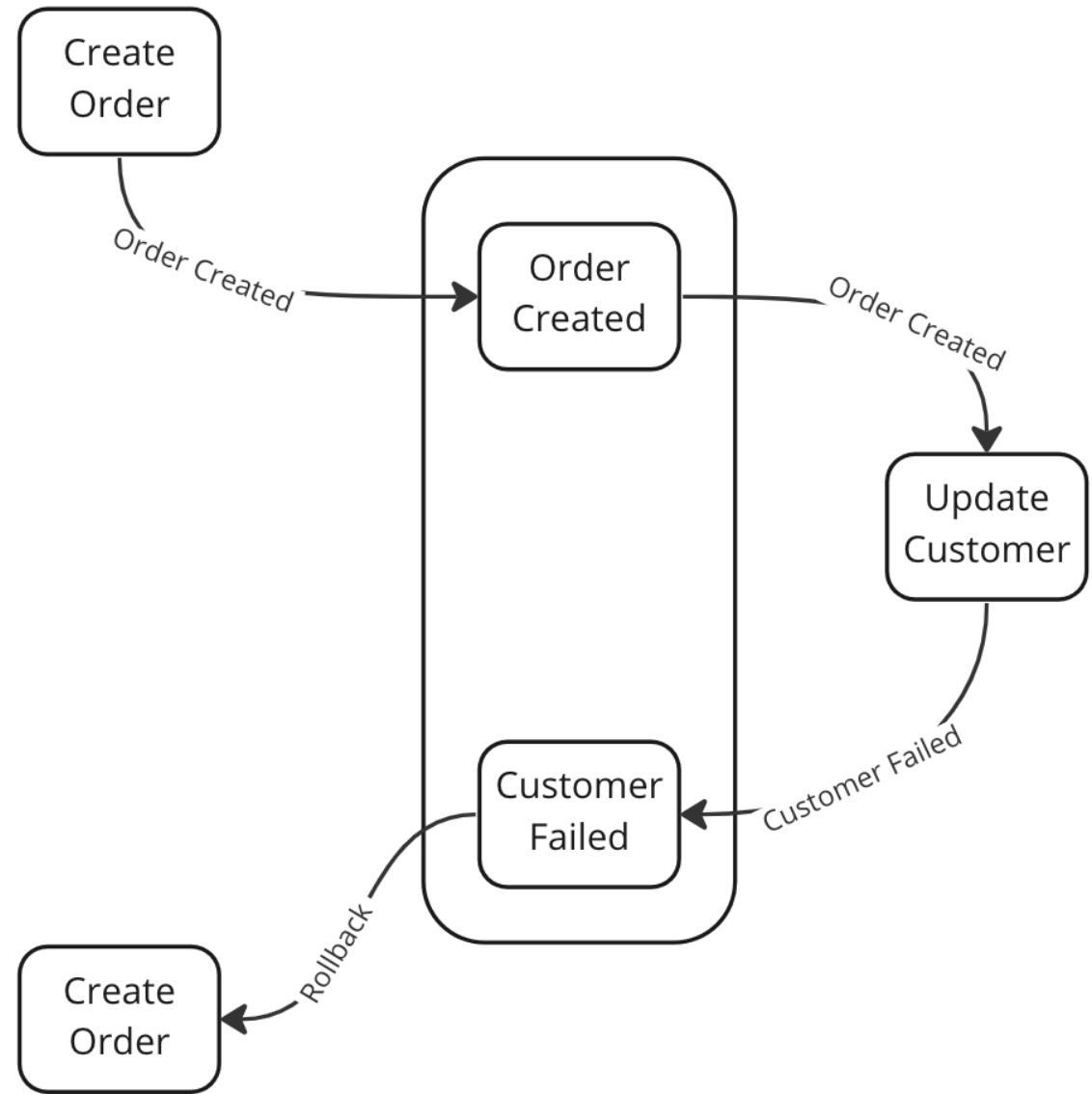


Saga

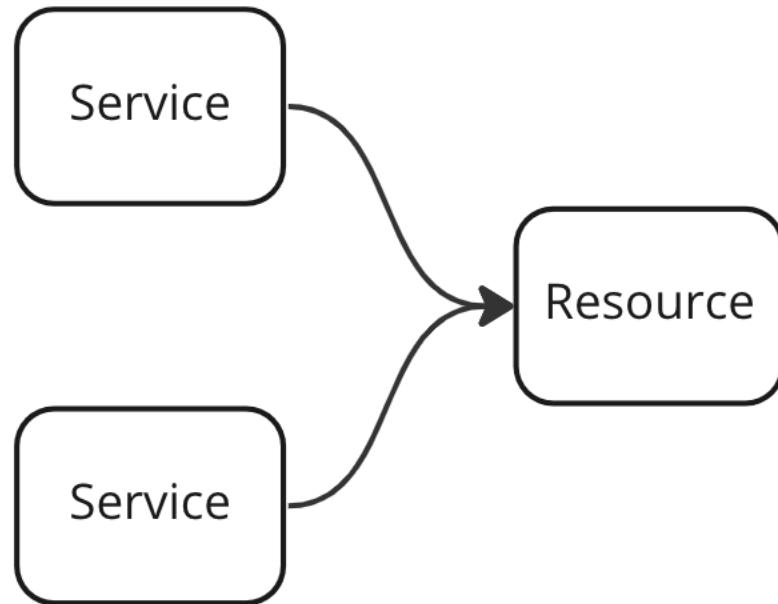
Представляет собой набор локальных транзакций.

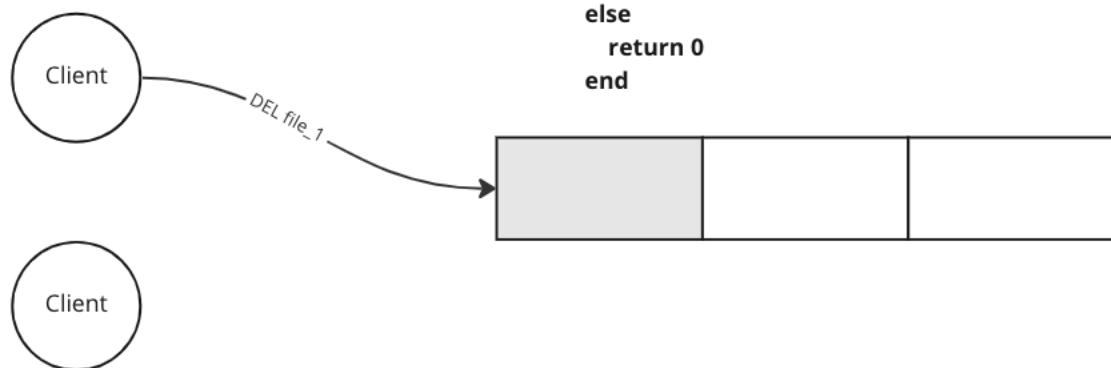
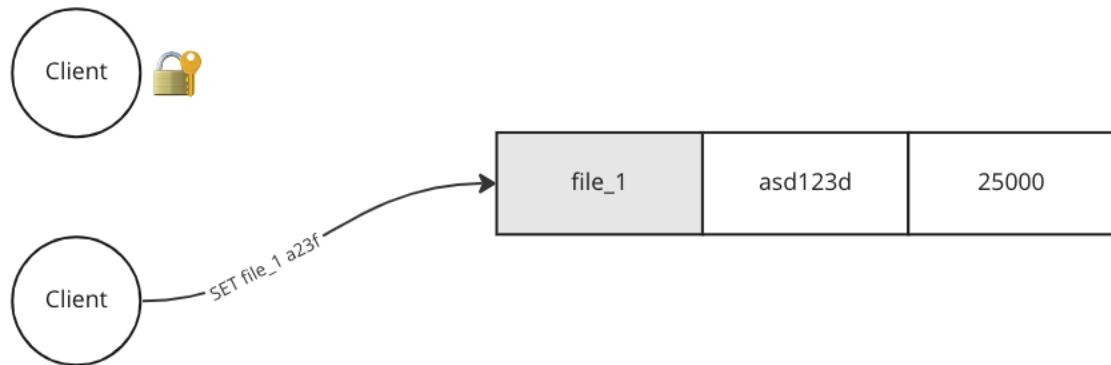
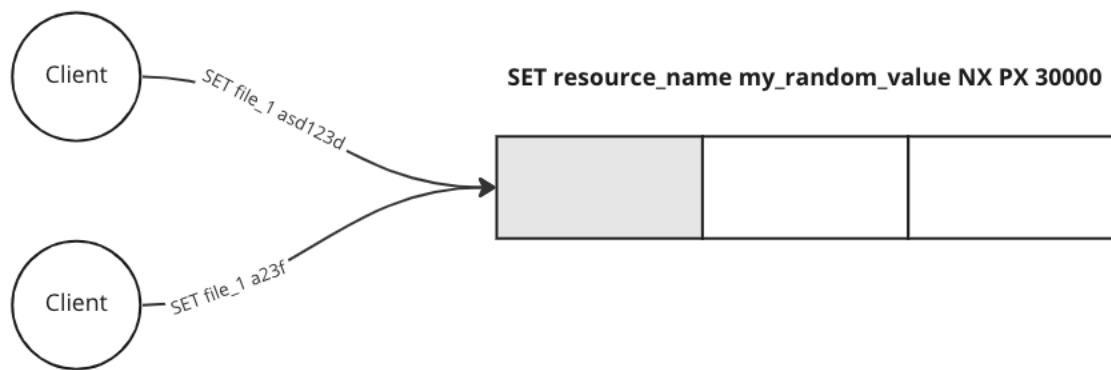
Каждая локальная транзакция обновляет базу данных
и публикует сообщение или событие, инициируя
следующую локальную транзакцию





Распределенные блокировки





Домашнее задание

- Пересмотреть занятие
- Для спроектированной баз(ы) добавить репликацию и шардирование (сообщения, посты, анкеты)
- Для спроектированной баз(ы) данных социальной сети добавить микросервисы
- Созвониться с партнером и обсудить вместе для каких еще систем вы бы использовали изученные паттерны (со звездочкой вместе попроектировать подсистемы)
- Погрузиться поподробнее в те темы, по которым осталось какое-то недопонимание
- Оставить отзыв о занятии в Google форме