CockroachDB

хипстерское поделие, или новая эпоха веба?

Виталий Левченко
Team lead @ mc2 software

Даниил Подольский
ex.CTO @ Git in Sky

Что будет

- CockroachDB
- Интрига
- Немного расследований

Почему мы?

Почему мы?

- Just because we can

Виталий Левченко

- Поддерживал ~200 шардов MySQL… Задолбался.
- Продвигаю современные технологий
- Люблю новое

Даниил Подольский

- Ищу надежную отказоустойчивую транзакционную СУБД
  - Без нее все очень трудно в мире обслуживания счетов
- Она должна быть быстрой

Что было раньше

- Шардинг MySQL / PostgreSQL

Что было раньше

- Шардинг MySQL / PostgreSQL
- Закрытые системы: Spanner, Dynamo etc

Что было раньше

- Шардинг MySQL / PostgreSQL
- Закрытые системы: Spanner, Dynamo etc
- Cassandra

Что было раньше

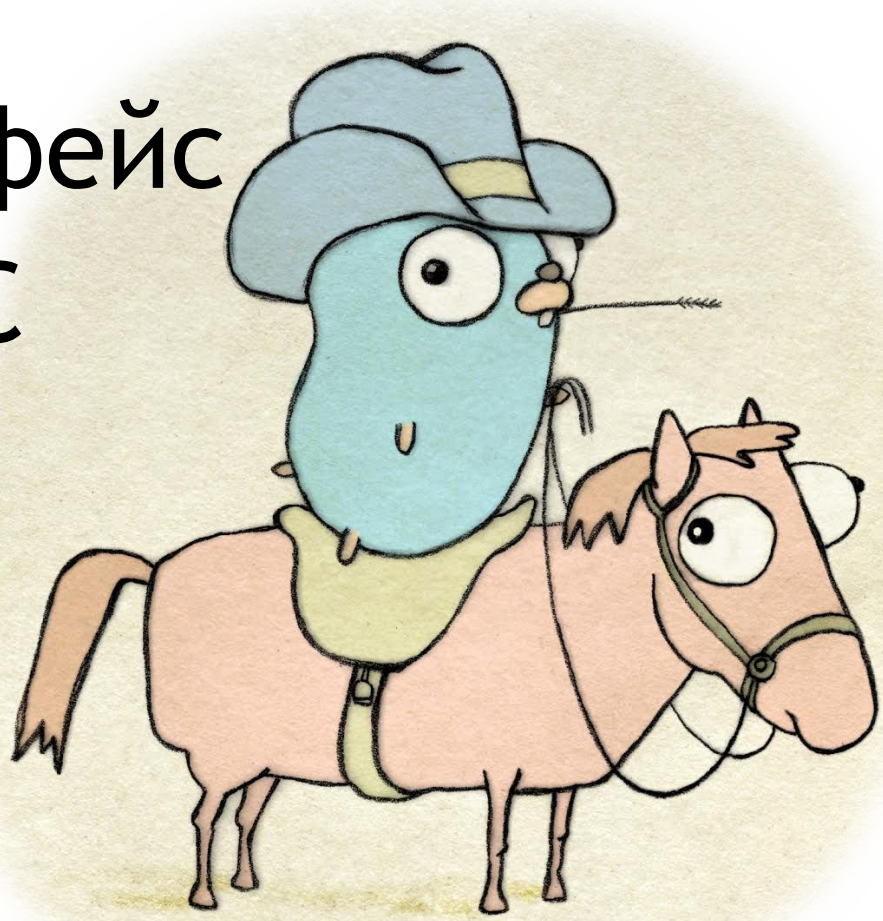
- Шардинг MySQL / PostgreSQL
- Закрытые системы: Spanner, Dynamo etc
- Cassandra
- MongoDB

CockroachDB

- Релиз 10 мая 2017
- Честная отказоустойчивость
- Прозрачный решардинг
- Распределённые транзакции

Под капотом

- Postgres интерфейс
- RocksDB + MVCC
- Raft
- Единый kv map



Отказоустойчивость и консистентность

- Serializable / Snapshot Serializable
- Конфигурируются в специальном range
- Range = 3 node Raft cluster ≤64Mb data
- Раскиданы по нодам
- Транзакции в них же
- Jepsen approved

Немного про Raft

- Sequential consistency
- Кворум 2N+1
- Устойчив к net split
- Научно доказан

Синхронизация времени

- Spanner: 7ms хватит всем!
- Мы можем так же:

```
env COCKROACH_LINEARIZABLE=true
```

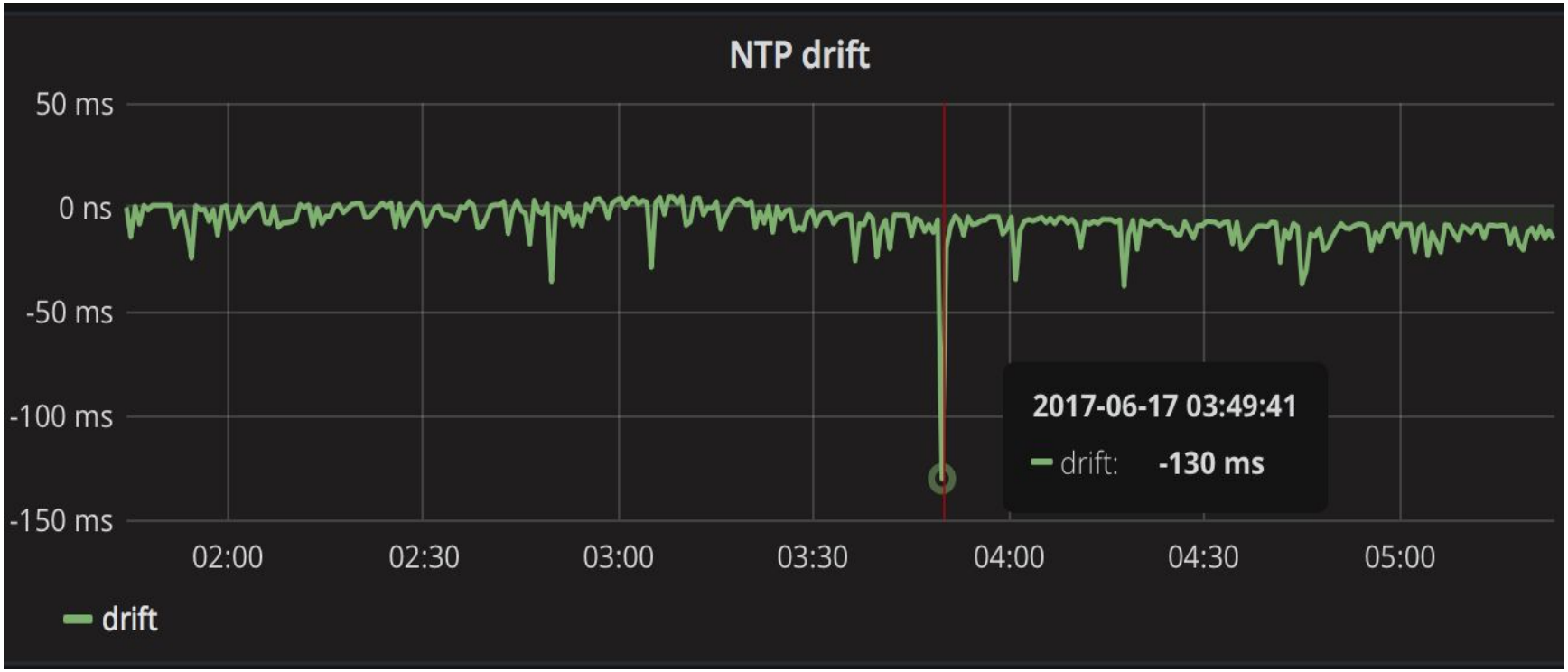
Синхронизация времени

- Spanner: 7ms хватит всем!
- Мы можем так же:

```
env COCKROACH_LINEARIZABLE=true
```

- GPS NTP server — дешёвый
- Но подводит железо

Синхронизация времени



Решардинг

- Автоматический split / join
- Распределяет по разным ДЦ

Геобалансировка

- Автоматический перенос leasehoder в ближний ДЦ
- Ручное указание split
- Автоматический split

Требования Роскомнадзора

149-ФЗ от 27.07.2006 «Об информации, информационных технологиях и о защите информации», ст.16, ч.4:

> Обладатель информации, оператор ИС … обязаны обеспечить:
>
> 7. нахождение на территории РФ баз данных информации, с использованием которых осуществляются сбор, запись, систематизация, накопление, хранение, уточнение (обновление, изменение), извлечение персональных данных граждан РФ.

Требования Роскомнадзора

- Может прокатить
- Лучше подождать до ноября

Боль админа

- Можете снять бекап
  - Но только за деньги
- сложности с внутренней сетью
  - Опять
- Средства сегментирования
  - недостаточно мощны

Тесты

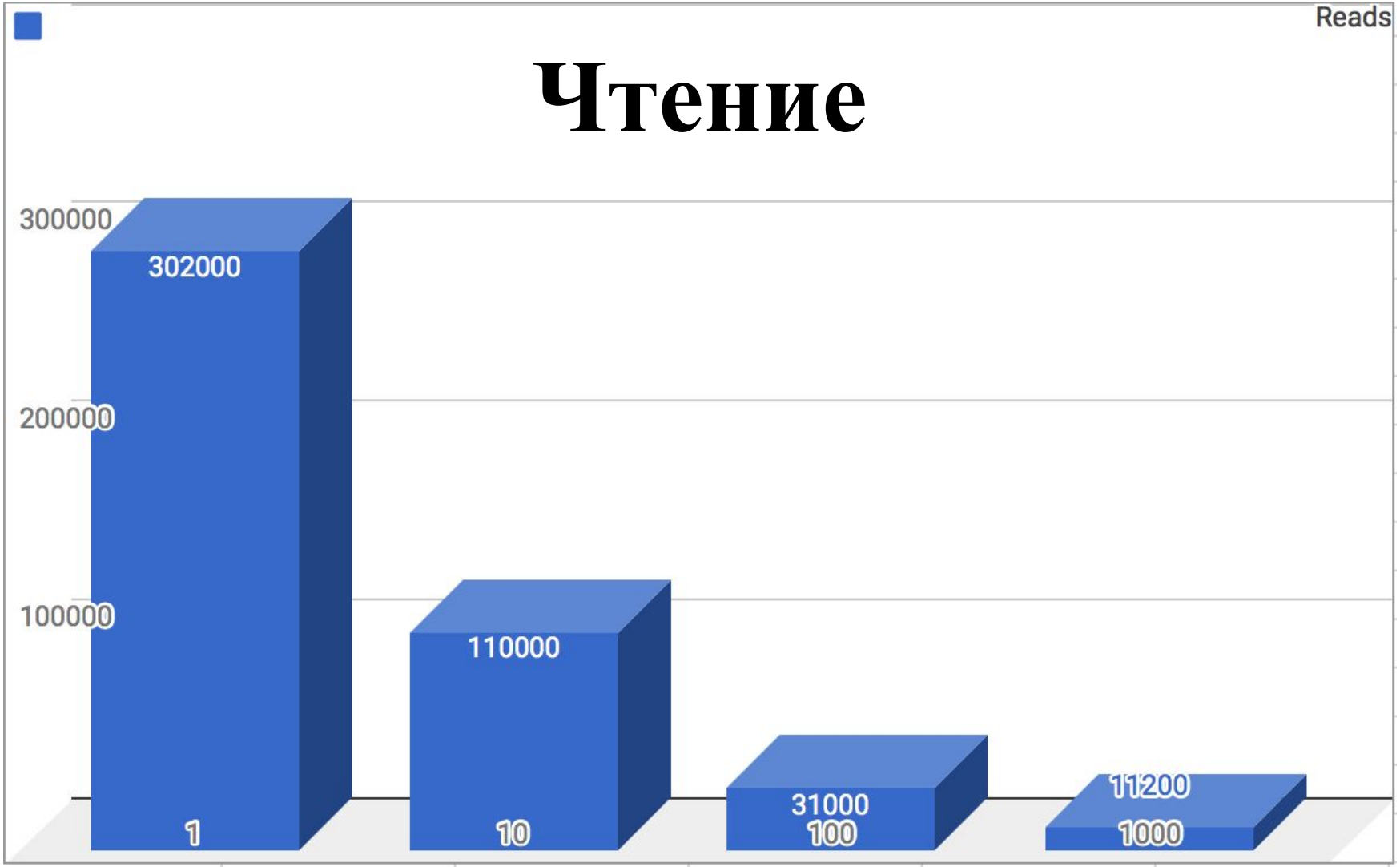
Тестовый стенд

- Dallas: Dell R220, 32Gb, SSD x9
- Ams: SameDO containers x3
- Ubuntu 16.04-2
- CockroachDB 1.0.2

Servers.com

- Решают нестандарные проблемы
- Обеспечили успех Prisma
- Удобные партнёры
- Хорошее железо
- Есть в РФ

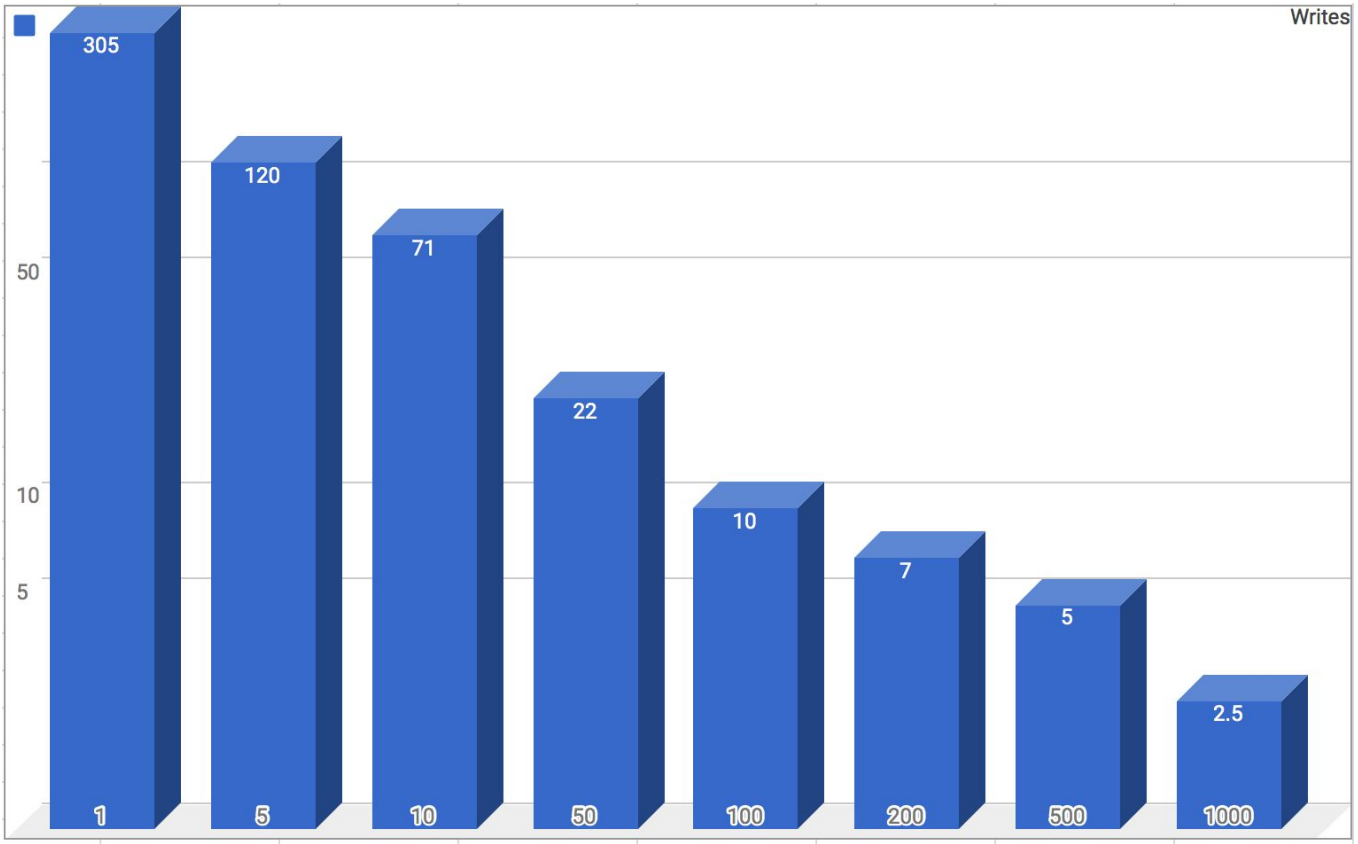
Чтение


Reads
300000
200000
100000
302000
1
110000
10
31000
100
11200
1000

Чтение

- Snapshot serializable

Запись



Запись

- Return nothing
- Updates = ins * 0.9
- Incr = 200 qps instead of 300
- On empty db x3

Join

- Interleaving = read
- Join 1 row = 500k

Аналитика

- ~ PostgreSQL
- 10s/50Gb data

Итого

- Очень быстрое чтение
- Медленная запись (временно)
- Упирается в io.
- Пригодно для аналитики

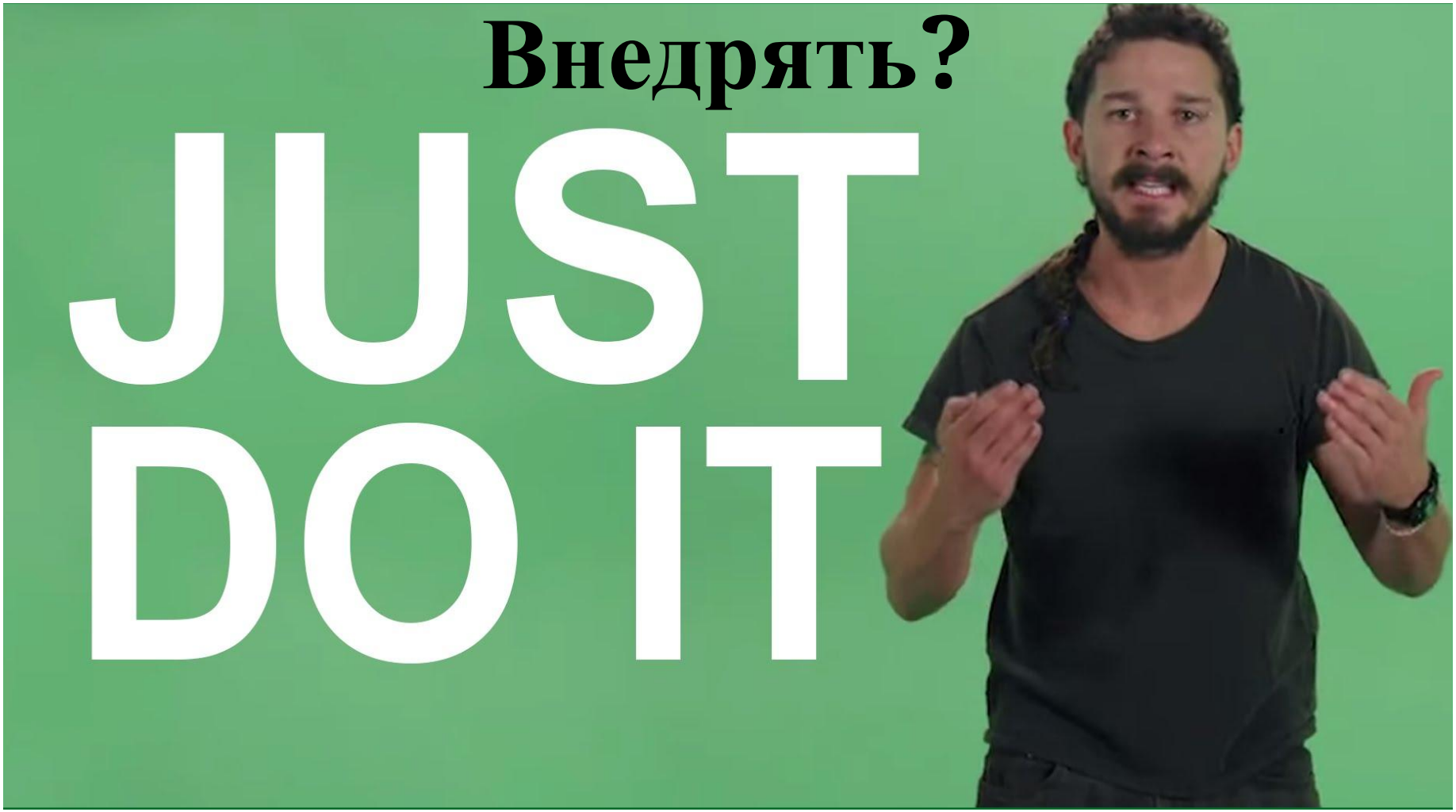
Внедрять?

- Если не жалко карьеры,
- Нужен Multi DC и хватает производительности — берите
- Особенно если есть быстрый сторадж в памяти

Внедрять?

- Внутри DC консервативные решения удобнее, быстрее и надёжнее (кроме Mongo)
- Лучше подождать до конца года

Внедрять?



Вопросы