Avito.ru
Объявления и Postgres

mtyurin@avito.ru // we are hiring

Тюрин Михаил

главный системный архитектор

Avito

Сайт бесплатных объявлений №1 в России *

Авто, недвижимость, работа, услуги, техника, одежда и многое другое. На сайте **31 641 046 объявлений.**

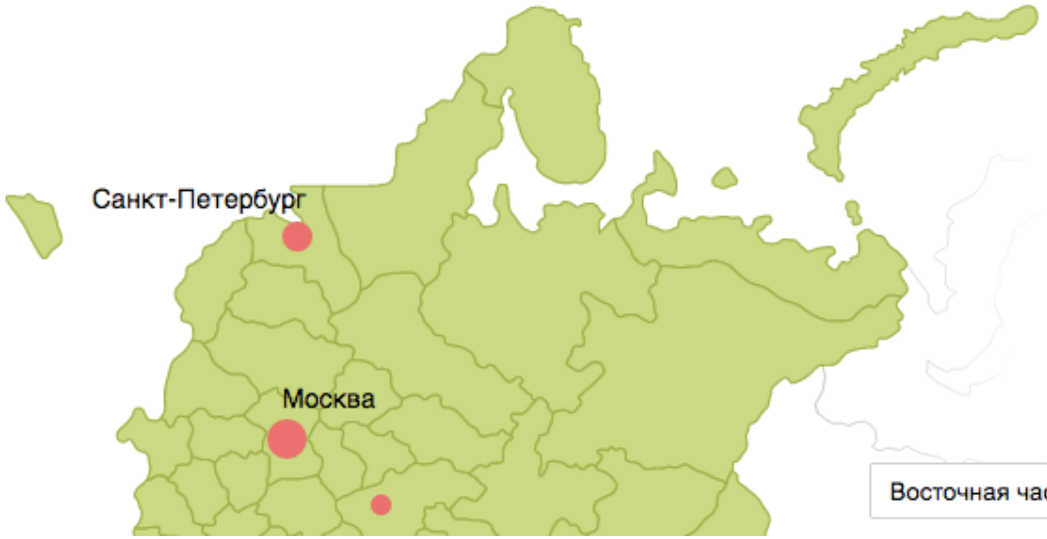
Личный кабинет ▾
Вход и регистрация

Подать объявление

Найдите объявление в своем городе | Свяжитесь с владельцем объявления | Совершите сделку!

АВТО | НЕДВИЖИМОСТЬ | РАБОТА | УСЛУГИ

Все категории



Москва

Санкт-Петербург
Волгоград
Екатеринбург
Казань
Краснодар
Нижний Новгород
Пермь
Ростов-на-Дону
Самара
Уфа
Челябинск

Адыгея
Архангельская обл.
Астраханская обл.
Башкортостан
Белгородская обл.
Брянская обл.
Владимирская обл.
Мордовия
Московская обл.
Мурманская обл.
Ненецкий АО
Нижегородская обл.
Новгородская обл.
Оренбургская обл.

Объявления Магазины Помощь ★ 0 Вход и регистрация

Avito Авто Недвижимость Работа Услуги ещё... Подать объявление

Любая категория | Поиск по объявлениям | Москва | Станция метро | Найти

искать только в названиях только с фото

Все объявления в Москве 6 274 998

Личные вещи 1 864 423 Для дома и дачи 674 297 Для бизнеса 171 149 Животные 81 308
Транспорт 1 763 094 Бытовая электроника 484 616 Услуги 150 209 NEW
Хобби и отдых 687 178 Недвижимость 297 934 Работа 100 790

Все | Частные | Компании По дате

☆ **Xeon Quad X5550 LGA 1366 3.06GHz Turbo 6.4GT/s**
2 000 руб.
Товары для компьютера
м. Калужская
Сегодня 00:53

☆ **Продавец цветов, выездная торговля**
30 000 руб.
Вакансии
ООО "Дорога цветов"
Сегодня 00:53

☆ **Weltmeister stella 3/4**
9 500 руб. ↘
Музыкальные инструменты
м. Тушинская
Сегодня 00:53

☆ **Картина 40х40 см. Холст, масло**

VIP-объявления

Монтаж и реконструкцию кровли
350 руб.
Частное лицо

hi!

avito.ru — «большой» мобайл и веб

- разработка с 2007-ого, активная фаза с 2010-го
- я, Михаил Тюрин, в компании с 2009-го (6 лет — похлопаем!)
- сейчас — я главный системный архитектор
- вырос из команды четырех первых третьих программистов
- сейчас большой департамент (на целый этаж)
- работаю с:
    - директором по разработке, девопс директором, dwh
    - тимлидами и ведущими разработчиками
    - командой баз данных (**Postgres**)

я в проекте практически с самого начала

- первая группа: php + свой_фрейморк + mysql
    - доисторические времена
- вторая группа: php + (postgres + tsearch2)
    - прототип почти был готов
- **третья группа:**
    - php + postgres
    - + Sphinx
    - и первый теле-трафик. теле. трафик!

я много всего знаю про то, как устроен Avito

- при чем тут postgres? — а с него всё началось!
  - первое, во что уперлись
  - но был опыт:
    - pg 7.4 — 8.2, **pg_upgrade** (formerly called pg_migrator)
    - sql/**plpgsq**
    - **hstore**
    - skytools: londiste (**pgq**), **plproxy** и walmgr,
    - php и memcached
    - **pgbouncer**
    - немного тюнить могли
    - знал по переписке Олега Бартунова

задавайте мне потом вопросы

- далее:
  - Sphinx
  - ну и как всё это индексировать?!
  - база начала расти
  - первая логическая репликация
    - skytools londiste
    - materialized view и deferred triggers
    - hstore
  - лучше научились тюнить
    - work_mem
    - «хинты» планеру
    - shared_buffers

много вопросов

- далее:
  - pgbouncer
  - начали писать свой класс db на php
  - первая очередь для писем
  - первый демон на php (skytools)
  - много кода на plpgsq

про связь postgres со всеми частями системы

- далее:
    - база растет
    - не успеваем индексировать
    - придумали параллельную индексацию
    - и материализованный снепшот (repeatable read)

    - и тут еще и очередь модерации

связь с поиском и вебом

- далее:
  - база растет
  - на логическую реплику переносим выдачу
  - появилось первое нормальное железо
    - psi raid bbu cache
    - raid10 15K hdd

про тюнинг всего и вся

- далее:
  - база растет
  - вводим еще один мастер — вертикальный шардинг
  - часть отчетов начинаем переносить на стендбай

- индекс бек-офиса — **все** объявления на Sphinx

про надежность и архив и восстановление

- далее:
  - база растет
  - появляется **хот** стендбай
  - часть выдачи переносим на хот стендбай
  - собственный архив (PITR)
  - отчеты считаем на отдельной машине из восстановленного бекапа
  - тюним дальше shared_buffers и checkpoint
  - начали тюнить linux

про fsync и tmpfs

- далее:
  - база растет
  - личный кабинет в tmpfs
  - в tmpfs
  - async commit
  - raid cache — write 146%

а какой у вас планировщик io и fs, почему?

- далее:
    - база растет
    - авито начал много зарабатывать
    - организуется dwh система (vertica $$$)
    - планируем варианты интеграции
        - отгрузка по времени
        - материализованные дельты
            - внутри мастера
            - и на логической реплики

а какие очереди и почему pgq?

- далее:
  - база растет
  - нужно развивать асинхронные механизмы работы с данными
  - асинхронные
  - pgq
  - xrpc
  - xrpcd — базы вызывают друг друга и php

какие размеры и нагрузки и чем мониторите?

- далее:
  - база растет
  - новое железо и $$D
  - ssd — хороши
    - быстрые ооочень
    - очень!
    - много iops
    - много место в мало юнитах
    - еще больше bbu cache — 2GB

какие сервера и сколько? а под базы?

- далее:
  - база растет и трафик всё это время тоже рос
  - и кол-во фич росло
  - ! но активная часть итемов не зависит от времени
  - геокодинг ходит асинхронно в Яндекс
  - не успеваем индексировать — отказываемся от снепшота, ставим репликацию на паузу перед индексацией, после возобновляем

как и где храните картинки, на дисках?

- далее:
  - база ростет
  - xdb: 16 нод кластер plproxy (16 баз, 8 физ машин)
  - + опять пригодился pgbouncer как роутер

как часто всё падает, где и почему?

- далее:
  - база и кол-во проектов растет
  - еще раз вертикально шардим второй мастер
    - кончилось место
    - через стендбаи
  - проект «хвоста»: продуктово ограничили ЖЦ объявления по времени

а как вы выкатываете код хранимок?

- далее:
  - база растет всегда — это с начала и очень на долго поражает
  - но потом привыкаешь и к этому
  - тюним вакуум и **bgwriter**
  - **реиндекс (concurrently)**

я опоздал, а почему не mysql?

- так где же всё таки лежат объявления (и деньги)
    - **postgres**
        - мастер базы
            - и личный кабинет в tmpfs
        - хот-стендбаи (и архив)
        - реплика сайта и реплика индексера
        - dwh лог (etl очередь)
        - xdb кластер
    - **sphinx**
        - идекс сайта
        - индекс бекофиса
    - redis, memcached
    - rabbitmq
    - tarantool
    - fluentd, mongodb
    - vertica
    - aaaaaaaaaaaaaaaaaaaaa

вы так много всего сделали, но всё равно теряете транзакци?!

- да!
- МЫ ТЕРЯЕМ НЕСКОЛЬКО ТРАНЗАКЦИЙ ДВА РАЗ В ГОД В СЛУЧАЕ АВАРИЙ

- потому что мы пока так решили для себя cap теорему

- в случае аварии переключаемся на резерв и
- !!! делаем процедуры восстановления
    - обобщенный UNDO лог в londiste
    - специфические процедуры для других связанных узлов

спасибо! вопросы!

- теперь можно пролистать всё быстро заново и задавать вопросы
- делиться впечатлениями
- наверняка я про самое интересное и не упомянул

http://www.slideshare.net/ssuserdc9298/pgconfru-2015-avito-postgresql
http://www.slideshare.net/ssuserdc9298/pgconfru-2015avito-recoveryv2

Миша
mtyurin@avito.ru
we are hiring

я часов до 5-ти тут
могу говорить много часов
следите за анонсами митапов в Авито