
Яндекс

Яндекс

A/Б-тестирование на больших объёмах

Михаил Буряков

Группа экспериментов

План

- Что такое А/Б тестирование
- Процедура заведения эксперимента
- Конфигурация экспериментов
- Инфраструктура

Пара слов об истории

Задача

Абстрактная задача

- Система с пользователями
- В неё вносятся изменения
- Нужно уметь измерить пользу этих изменений

Пример

- Веб-сервис по сравнению пары фотографий котов
- Пользователь выбирает наиболее понравившуюся
- Варианты выбора: левый, правый, против всех
- Хотим обоснованно улучшать сервис

Решение

Абстрактное решение

1. Придумаем гипотезу
2. Запустим на две группы пользователей разные варианты
3. Посчитаем показатели качества
4. Сравним отличия с помощью статистических критериев

Особенности

Особенности для больших объемов

- Нет непосредственного контакта
- Миллионы наблюдений
- Шум

Вспомогательные термины

Экспериментальная выборка

- Подмножество пользователей, на которых срабатывает конкретное изменение
- Набор параметров (флагов) включающих экспериментальную функциональность

Эксперимент

- Совокупность из нескольких выборок
- Обычно одна из них контрольная (без изменений)
- Остальные включают какие-то изменения

Срез данных

- Подмножество данных полученное по указанному условию
- Пример: запросы с определённым регионом, языком, коммерческие и т.п.

От абстрактных котов к реальной жизни

ОсложненияЯ

В случае Яндекса, осложняют задачу экспериментирования следующие факты:

- Много сервисов
- Много экспериментов и экспериментаторов
- Много различных направлений для улучшений в каждом сервисе

ОсложнениЯ

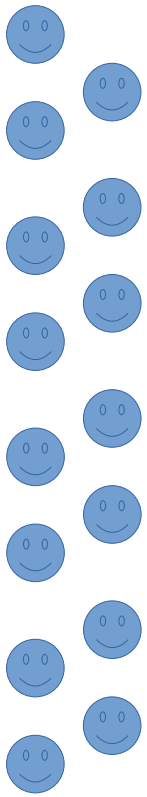
В случае Яндекса, осложняют задачу экспериментирования следующие факты:

- Много сервисов
- Много экспериментов и экспериментаторов
- Много различных направлений для улучшений в каждом сервисе
- За 2017 год: >5000 экспериментов, >400 единовременно

Система и порядок

- Общая инфраструктура и инструменты
- Процедура проведения экспериментов

Эксперименты в Яндексе



Пользователи

Запросы

Изменённая выдача

Яндекс

Инфраструктура экспериментов

Заявки

Результаты

Разработчики
Менеджеры
Аналитики

Типы экспериментов

- Часть сервиса
- Сервис
- Кросс-сервисные

- Улучшения (потенциальные)
- Ухудшения
- AA-тесты (+ сборы данных)

Процедура

Шаги проведения эксперимента

1. Подготовка
    - Заведение заявки
    - Тестирование
    - Одобрение
2. Проведение
    - Очередь
    - Включение
    - Сбор информации
3. Завершение
    - Анализ результатов
    - Принятие решения

Заявка

Заведение заявки

- Цель
- Критерий выкатки
- Тип: интерфейсы, ранжирование и ...
- Флаги для A и B
- Целевые срезы для расчёта метрик
- Ограничения (страны, платформы, браузеры и ...)

Заведение заявки - пример

- Цель: улучшить верстку результатов поиска
- Критерий выкатки: улучшение основных интерфейсных метрик
- Тип: интерфейсы
- Флаги для A и B: A: {}, B: {“good_interface”: true}
- Целевые срезы для расчёта метрик: desktop, touch, коммерческие запросы, частотные запросы, редкие запросы
- Ограничения: 10%, Россия

Тестирование

Цели тестирования

- Ничего не падает
- Флаги корректно срабатывают

Способы тестирования

- Вручную прощелкать в браузере
- Автоматическое
  - Проверки работоспособности модулей сервисов
  - Асессорские оценки

Одобрение

Одобрение

- Проблема: как не включать некорректные эксперименты

Одобрение

- Проблема: как не включать некорректные эксперименты
- Решение: модерация заявок экспертами
  - Понятность описания
  - Корректность флагов
  - Необходимость дополнительных тестов
  - Советы

Ожидание

Очереди

- Проблема:
  - Заявок много
  - Пользователей ограниченное количество
- Следствие:
  - Образуется очередь

Как ускорить нахождение в очереди?

- Одномерная схема: 1 пользователь и 1 эксперимент
- Многомерная схема: 1 пользователь и несколько экспериментов

Как ускорить нахождение в очереди?

- Одномерная схема: 1 пользователь и 1 эксперимент
- Многомерная схема: 1 пользователь и несколько экспериментов
- ~50 измерений

Запуск

Сбор и выкатка конфигурации

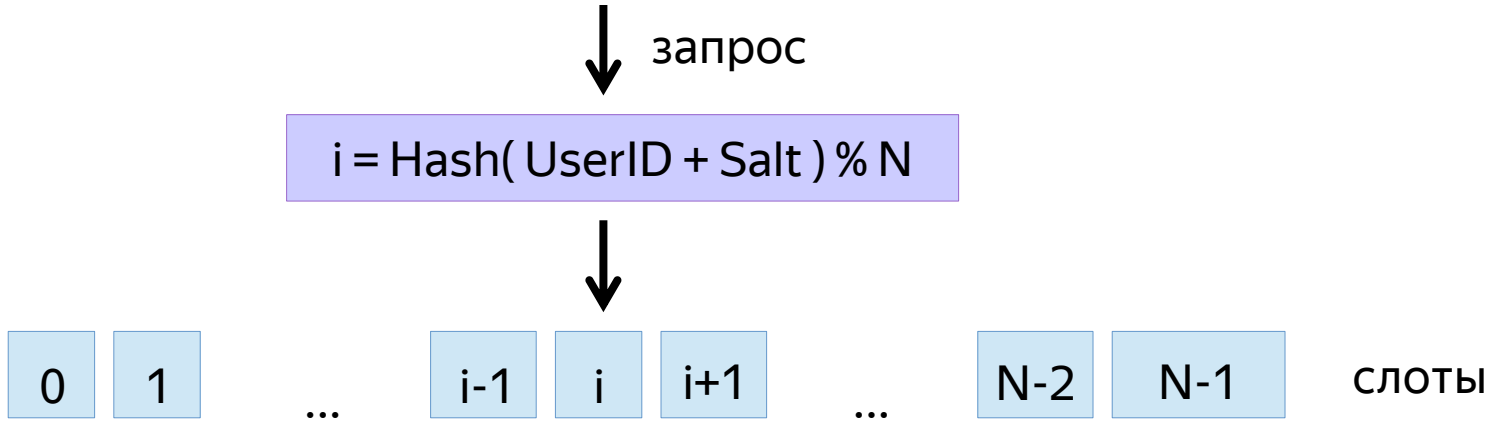
- Осуществляется релиз-менеджером (во многом автоматизирован)
- Сбор конфигурации: удаление просроченных и добавление новых
- Выкатка конфигурации несколько раз в день

Конфигурация экспериментов

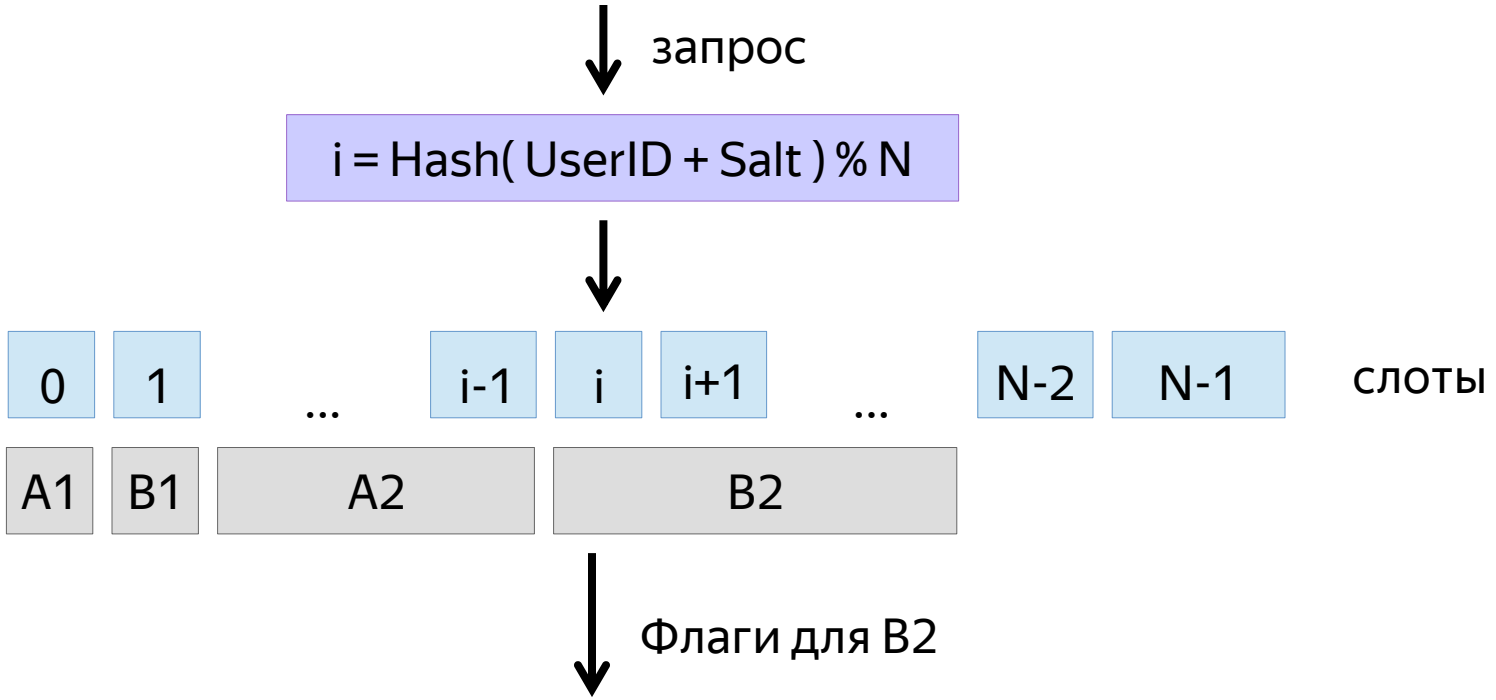
Правила разбиения на эксперименты

- F(request) = { мн-во экспериментов }
- Граф решений по:
  - User ID
  - Query
  - URL
  - Region
  - User-Agent
  - Timestamp
  - ...

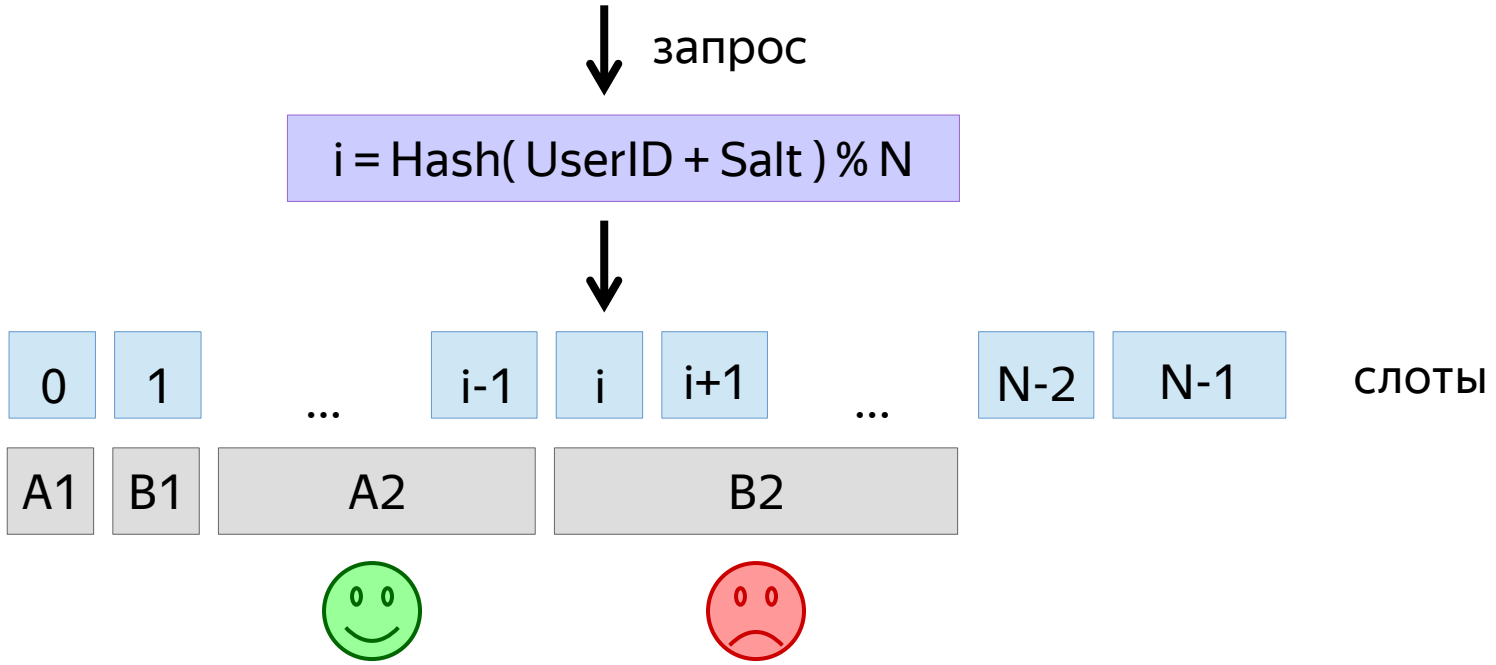
Разбиение



Разбиение

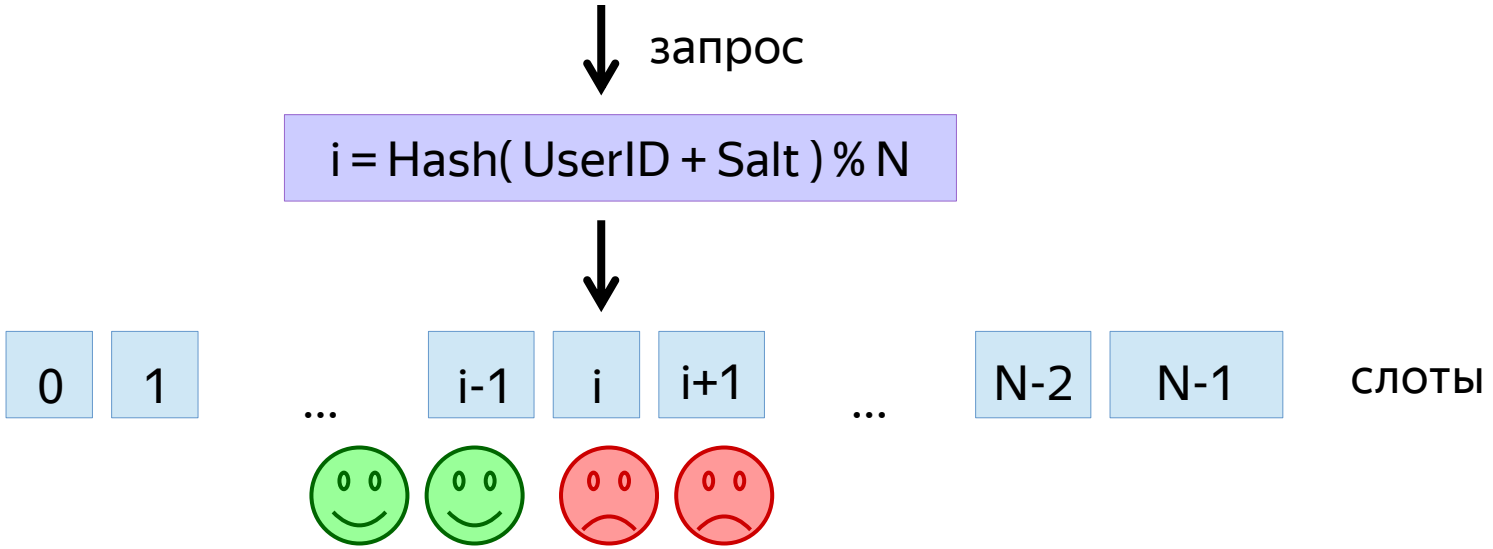

запрос
i = Hash(UserID + Salt) % N
0
1
...
i-1
i
i+1
...
N-2
N-1
слоты
A1
B1
A2
B2
Флаги для B2

Проблема: память пользователей



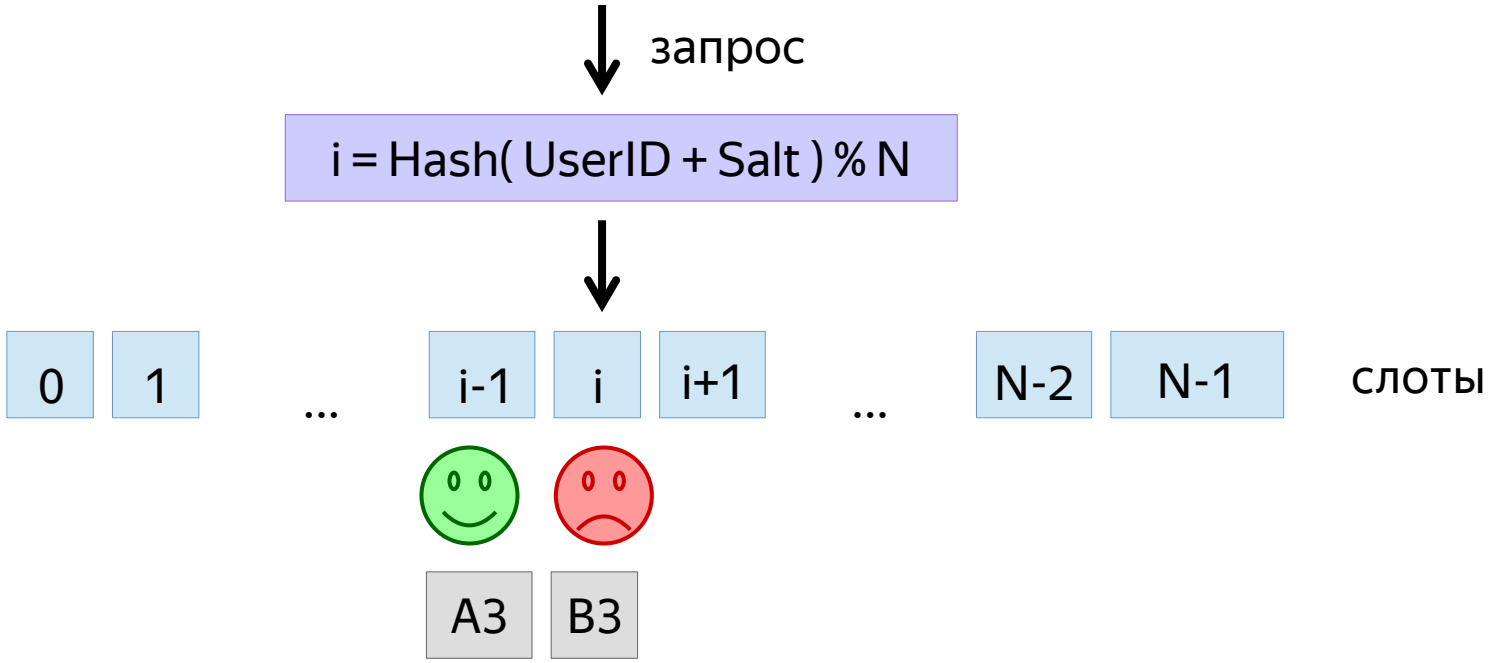
Пользователи привыкают

Проблема: память пользователей



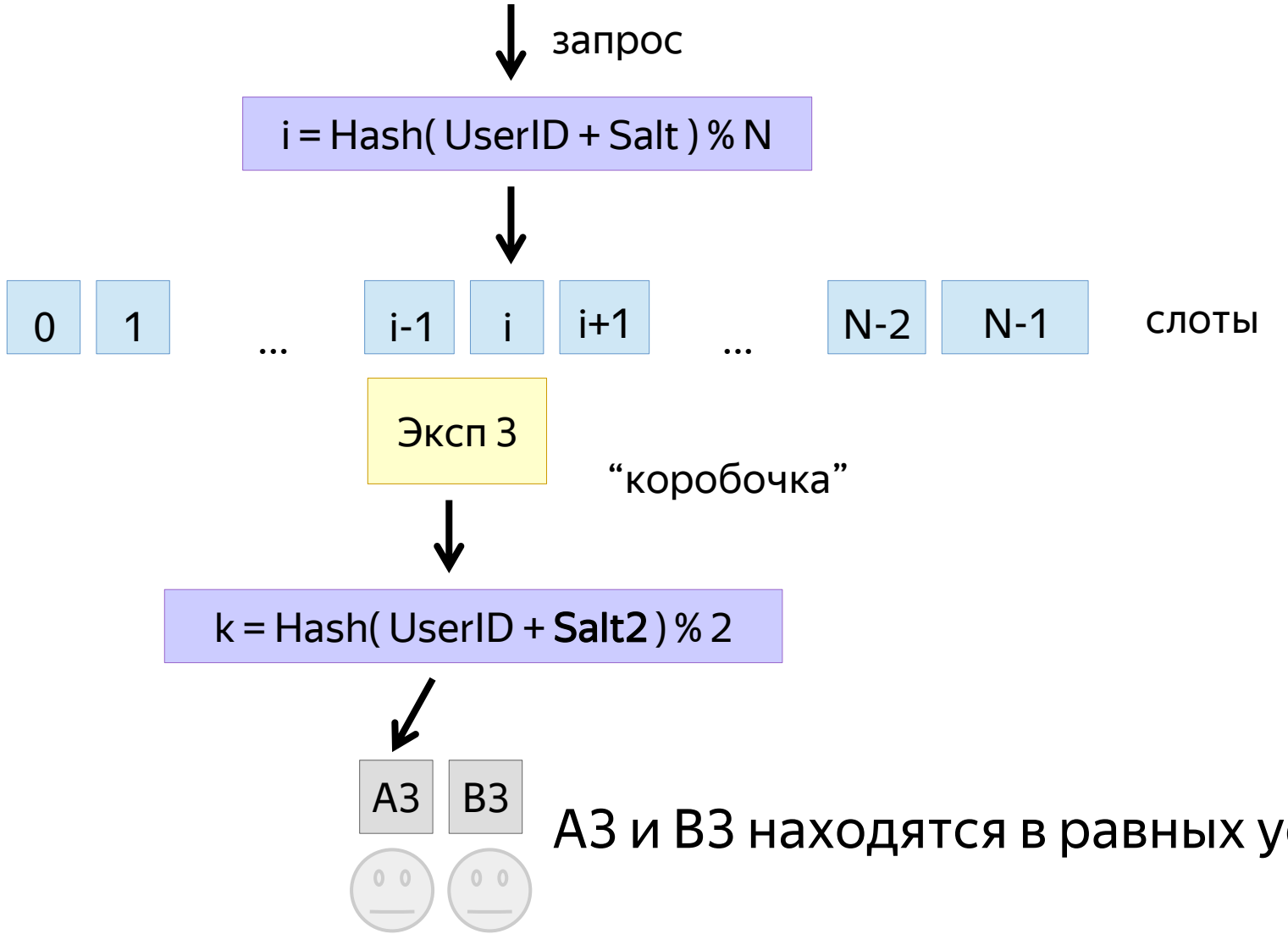
Даже после завершения эксперимента пользователи ведут себя по-разному

Проблема: память пользователей



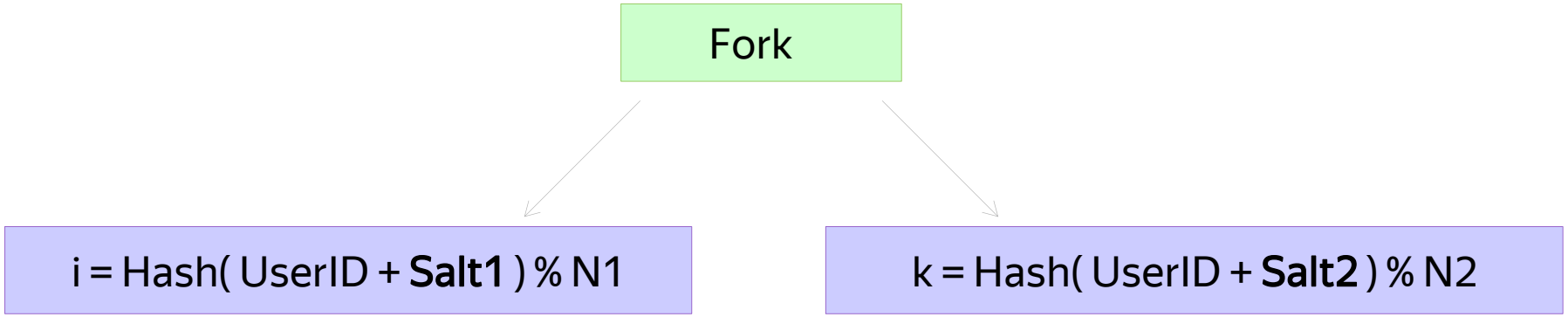
A3 и B3 находятся в неравных условиях

Решение: дополнительное перемешивание



A3 и B3 находятся в равных условиях

Многомерная схема



Пользователь попадает в два эксперимента одновременно

Сбор конфигурации

Проблема: сбор конфигурации

Как разместить?

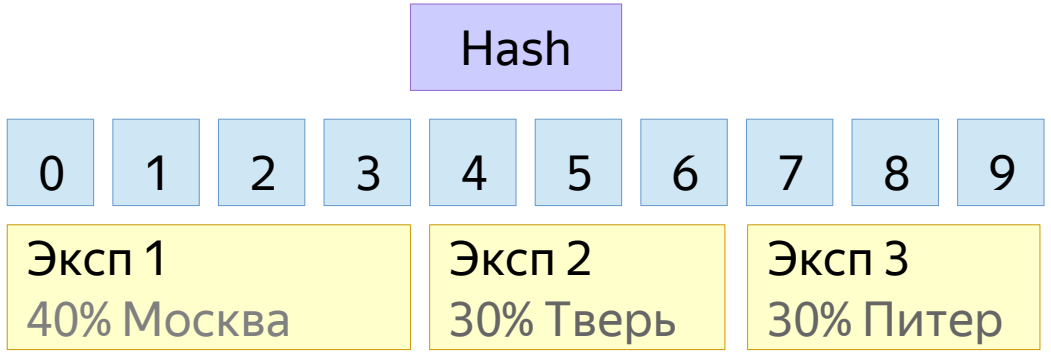
Эксп 1
40% Москва

Эксп 2
30% Тверь

Эксп 3
30% Питер

Эксп 4
60% Россия

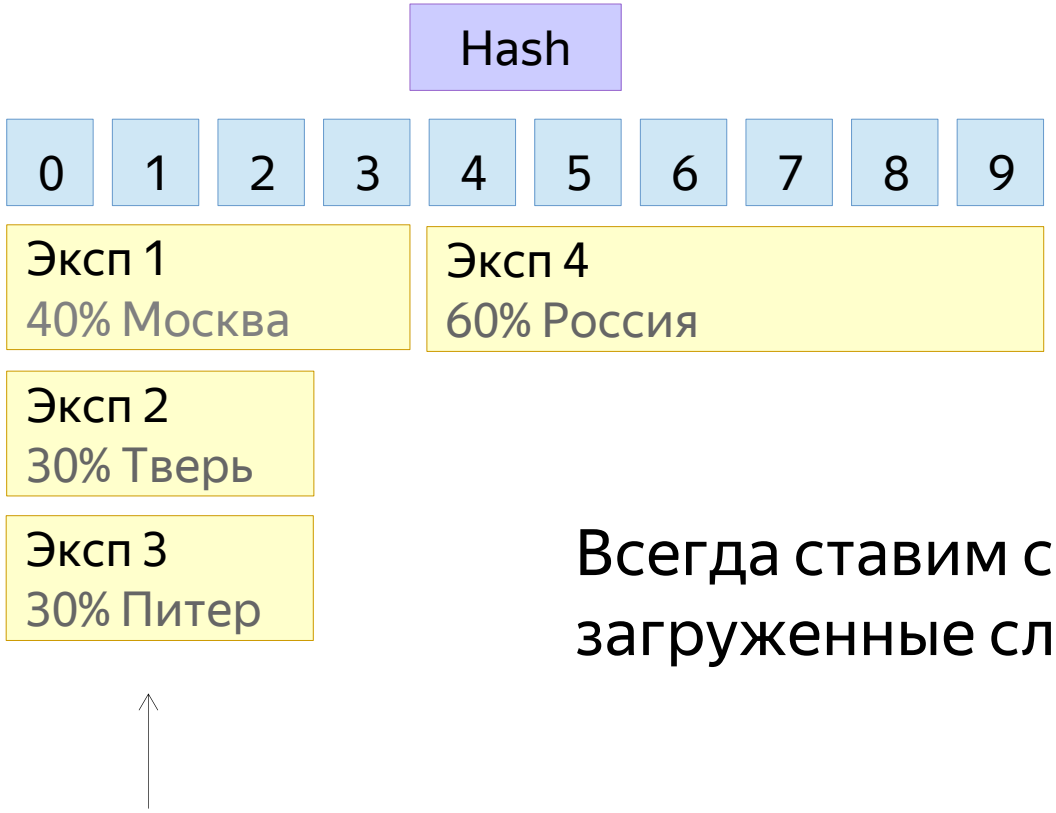
Проблема: сбор конфигурации



Не поместился!

Эксп 4
60% Россия

Проблема: сбор конфигурации



Всегда ставим сначала на наиболее загруженные слоты

На одни и те же слоты

т.к. не пересекаются по региону

Как обрабатывается конфигурация

Физика процесса

- Разбиение происходит на самом верху обработки запроса
- Конфигурация – файл на балансерах
- Как можно более простой процесс обработки
- Pull-процесс раскладки
- Экспериментальные флаги – HTTP-заголовки

Мониторинг

Мониторинг

Новая конфигурация ничего не сломает?

- Мониторятся ключевые показатели поиска

Как найти плохой эксперимент?

- Графики по основным метрикам
- Быстрое автоматические обнаружение аномалий
- Экстренное отключение

Что считать

Вкусы и числа

- Решить, что значит «хорошо»
- Метрика как показатель направления хорошести
- Улучшение метрик как стремление к идеалу

Метрики

- Кол-во пользователей, запросов и кликов
- Доля некликнутых
- CTR-ы
- ... и тысячи других

Метрики

- Кол-во пользователей, запросов и кликов
- Доля некликнутых
- CTR-ы
- … и тысячи других (>4000 различных метрик)

Нужны данные

Логирование пользовательских действий

- Выделим возможные действия пользователя
- Сохраняем на сервере
- Агрегируем и складываем в хранилище

Зоопарк данных

- Разный формат
- Разный размер
- Разное время доставки
- Распределённая ответственность

Вопросы доставки и агрегации

- Сложная инфраструктура и отлаженные процессы
- Совместимые форматы данных
- Общая библиотека для работы с логами
- Хранение для дальнейшей обработки

Расчёты

Расчёт метрик по экспериментам

- Распределённая обработка
- Построение выжимок
- Хранение для быстрого доступа

Анализ результатов

Просмотр метрик

- Интерфейс просмотра метрик
- По указанным промежуткам дней
- В нужных экспериментах и срезах

Исследование на срезах

- Нужно изучить совокупности срезов
- Интересно "повертеть" данные
- Поможет инструмент для быстрого анализа по произвольным совокупностям указанных срезов

Принятие решения

Обоснованность – статистические тесты

- Mann-Whitney
- T-test

Выкатываем или нет

- Проверка критерия выкатки
- Продуктовые соображения
- Одобрение экспертов
- ~28% экспериментов принимаются

Датасет

- История принятия решений
- Данные для исследований
- Поиск по истории экспериментов

Вопросы?

Спасибо за внимание!

Михаил Буряков
Группа экспериментов

 buryakov@yandex-team.ru