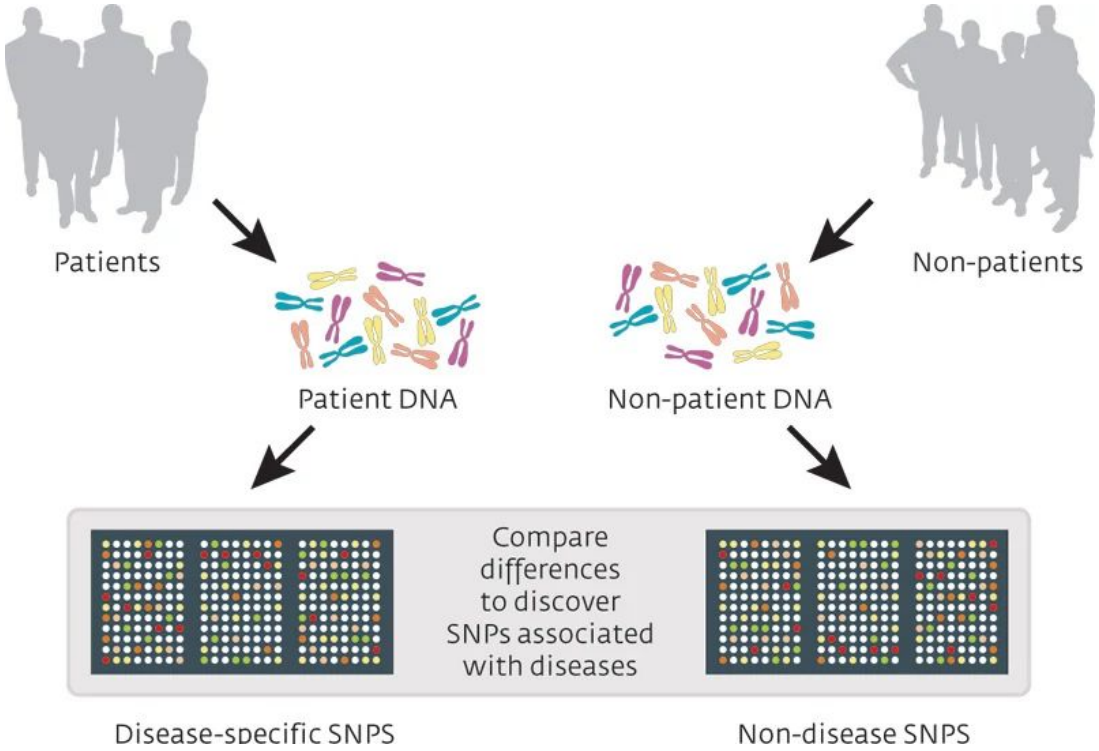
ClickHouse приходит в генетику

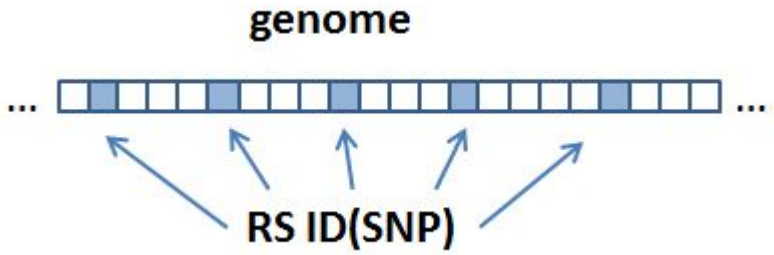
Про что расскажем

1. Как получаются генетически данные?
2. Как они выглядят? И зачем их хранить вместе?
3. Похожи ли на метрику?
4. Как хранили раньше?
5. А сколько всего данных?
6. Быстро работает? Есть статистика?
7. Что если положил не то?
8. А если не нужна вся база?
9. Итого

Как проводят полногеномные исследования ассоциаций(GWAS)?



Что в итоге? GWAS?



Сколько всего уникальных позиций?

SNP - место на геноме, в котором у разных людей стоят разные буквы. Имеет идентификатор (**RS ID**) и координаты (**Chromosome, Position**). Во всех остальных местах буквы одинаковые (мало людей участвует в исследовании)

Интересный gwas?

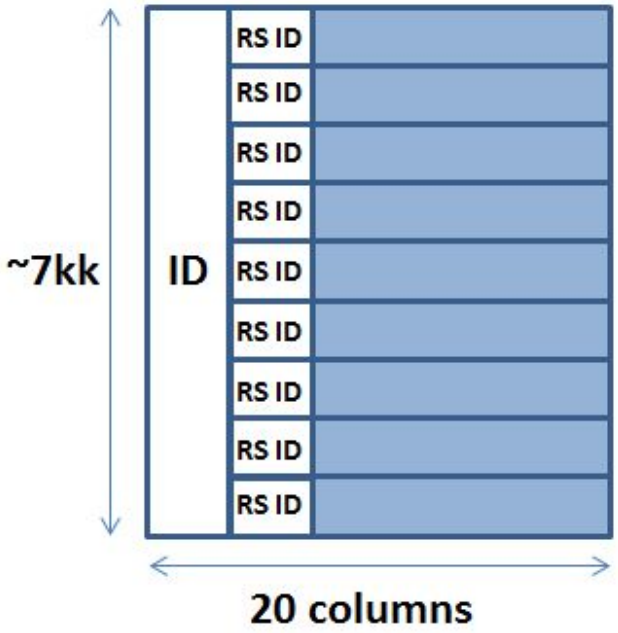
Значимость - исследование делают против признака. Насколько замена буквы в SNP влияет на признак

Word-to-vec на генетике?

Совместность - доступ ко многим GWAS позволяет исследовать перекрёстные зависимости между признаками

Как выглядят данные?

GWAS



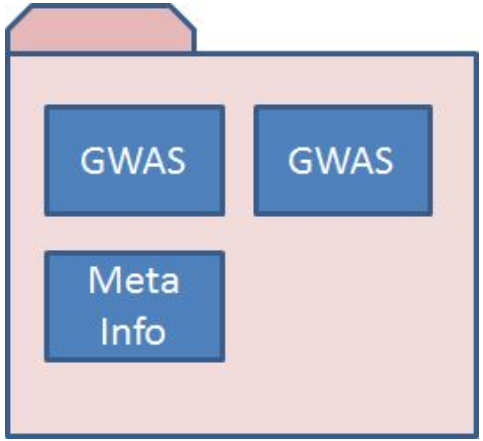
Метрика?

- **ID** - идентификатор сайта
- **RS ID** - идентификатор посетителя
- **Столбцы** - агрегированные метрики пользователя по сайту (среднее время на сайте)

Или нет?

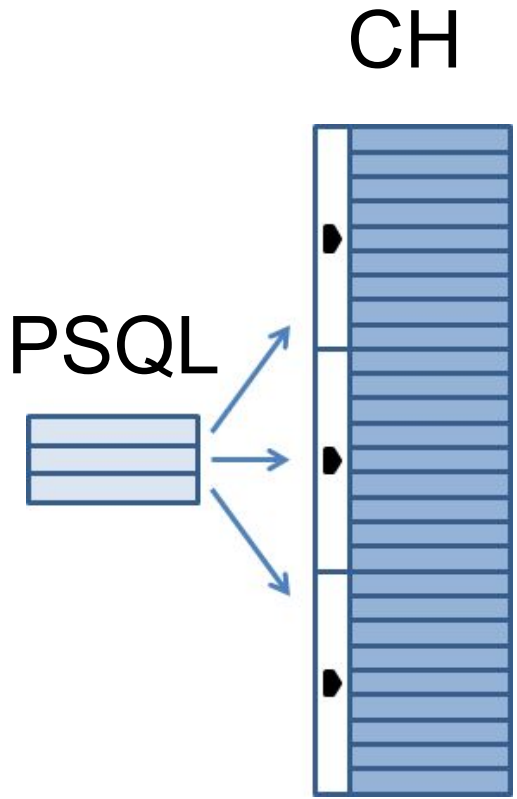
- Пишем не постоянно
- Загружаем данные блоками
- Не модифицируем

Как хранили раньше?



- Распределённость?
- Параллельные запросы?
- Агрегирующие запросы?

Как храним мы?



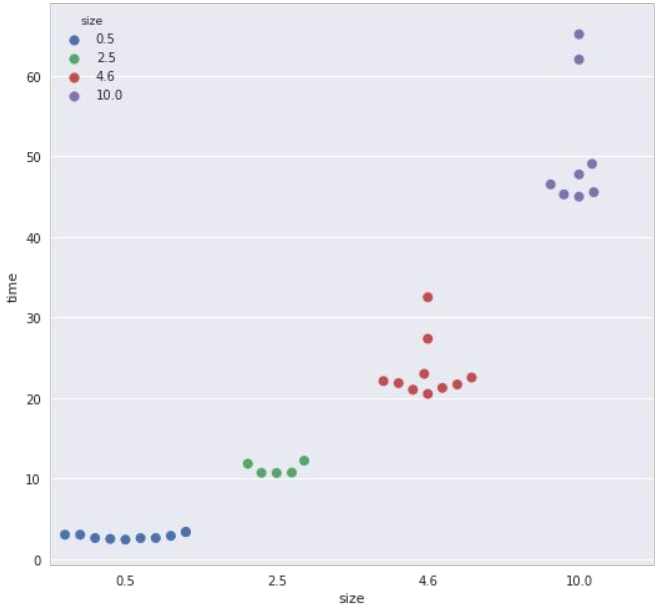
Сколько данных?

3.5 ТБ в формате csv -> 1.4 ТБ в ClickHouse
~ 3700 GWAS
~ 7 миллионов записей(SNP) для одного GWAS
~ 27 миллиардов записей

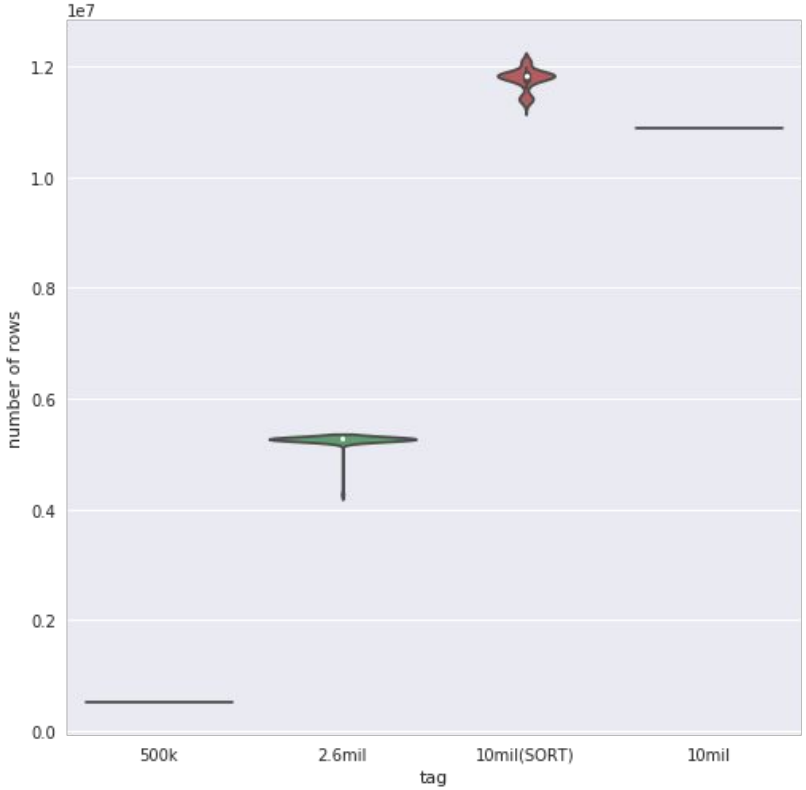
То, что вы боялись проверить:
Включение занимает 1.5 часа
Возможно ускорить?

Быстро работает?

Анализы написаны на Python, поэтому замеры производительности на Python



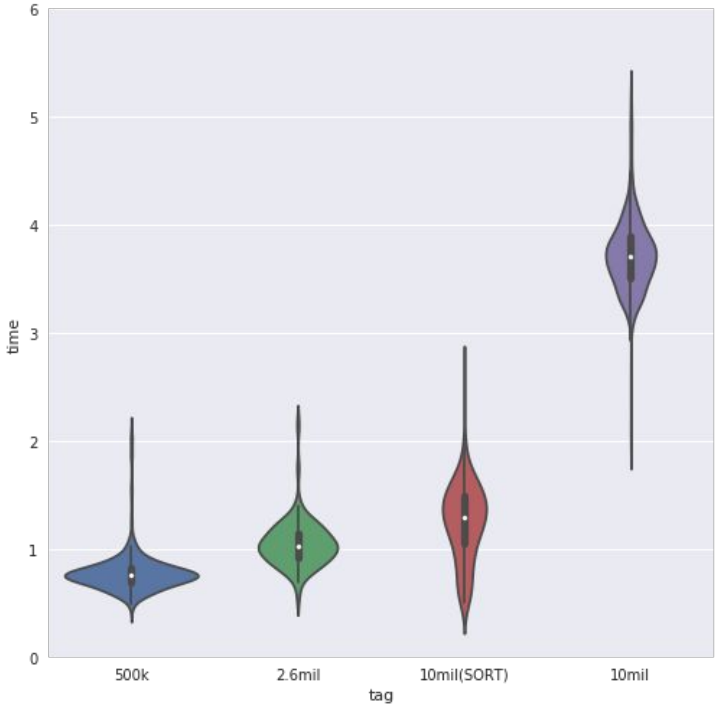
Время(секунды) выгрузки всего GWAS в зависимости от размера(миллионы)



Группы GWAS в системе по размеру (десятки миллионов)

А профильные запросы?

“Из GWAS с номером N дай все SNP в первой хромосоме между такими-то координатами”



Время(секунды) выполнения запроса против размера GWAS

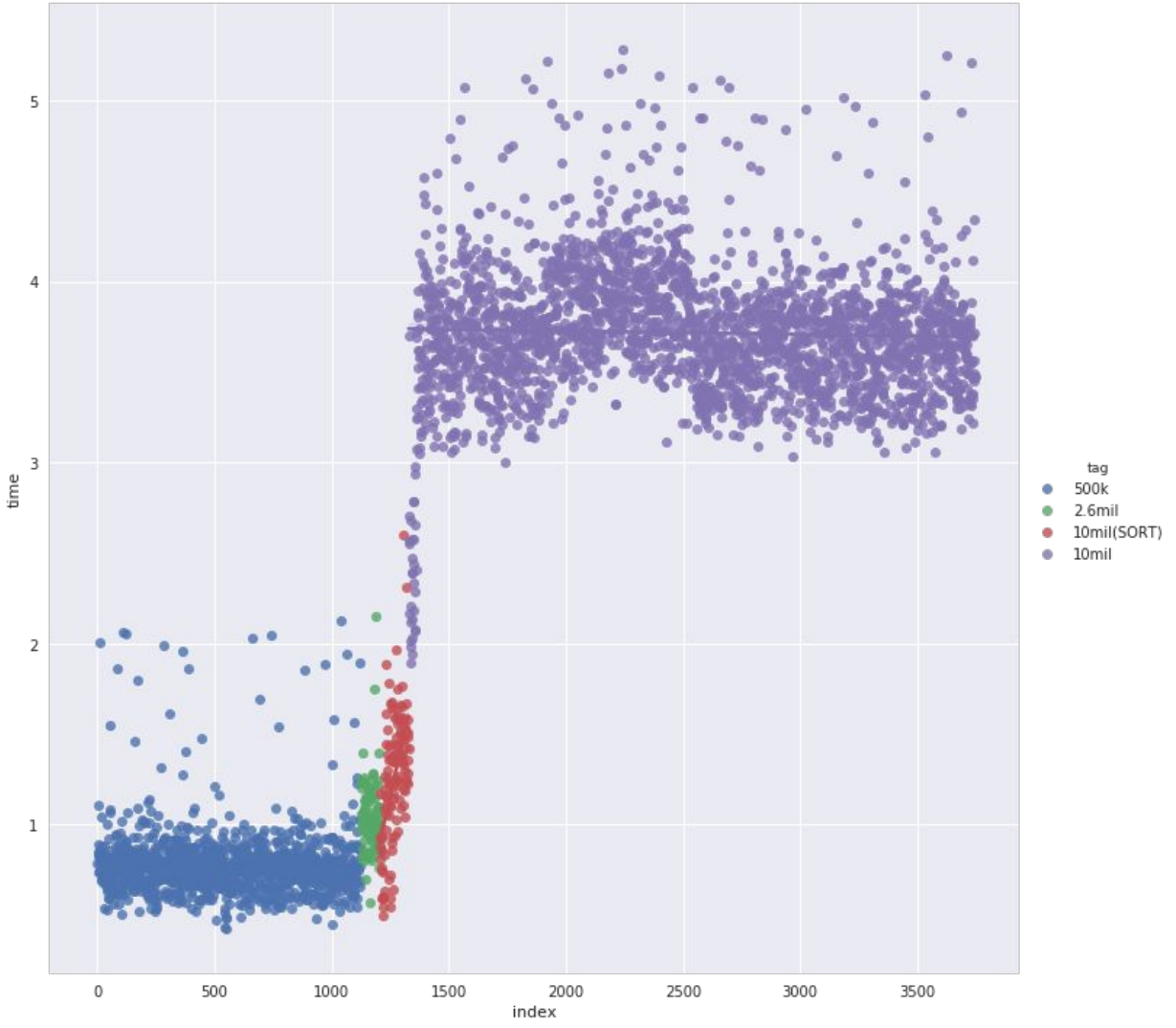
Почему большая дисперсия для каждой группы?
Откуда разница для равноразмерных GWAS? Размер выгрузки?

Время(секунды) выполнения запроса против порядкового номера GWAS

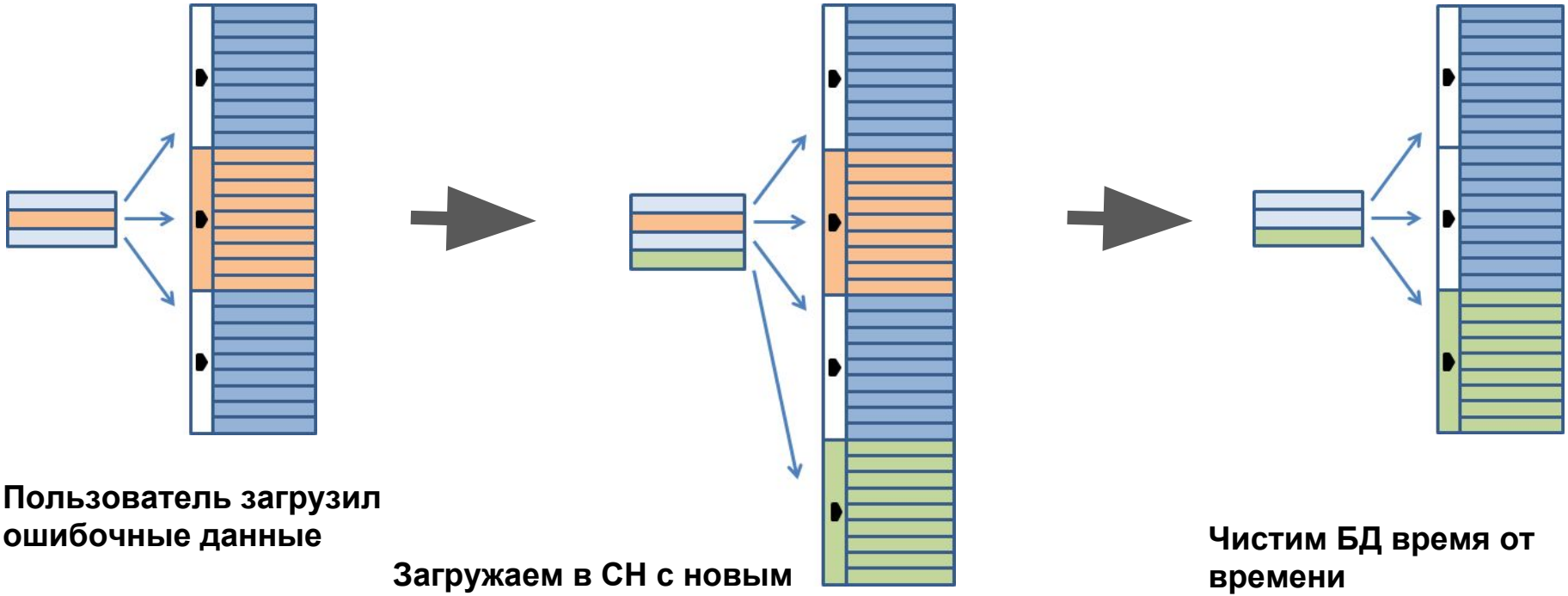
Почему быстро на первых GWAS?
Почему равномерные выбросы?

Нет зависимости скорости от положения GWAS

Время(секунды) выполнения запроса для разных групп GWAS



Что если загрузил не то?



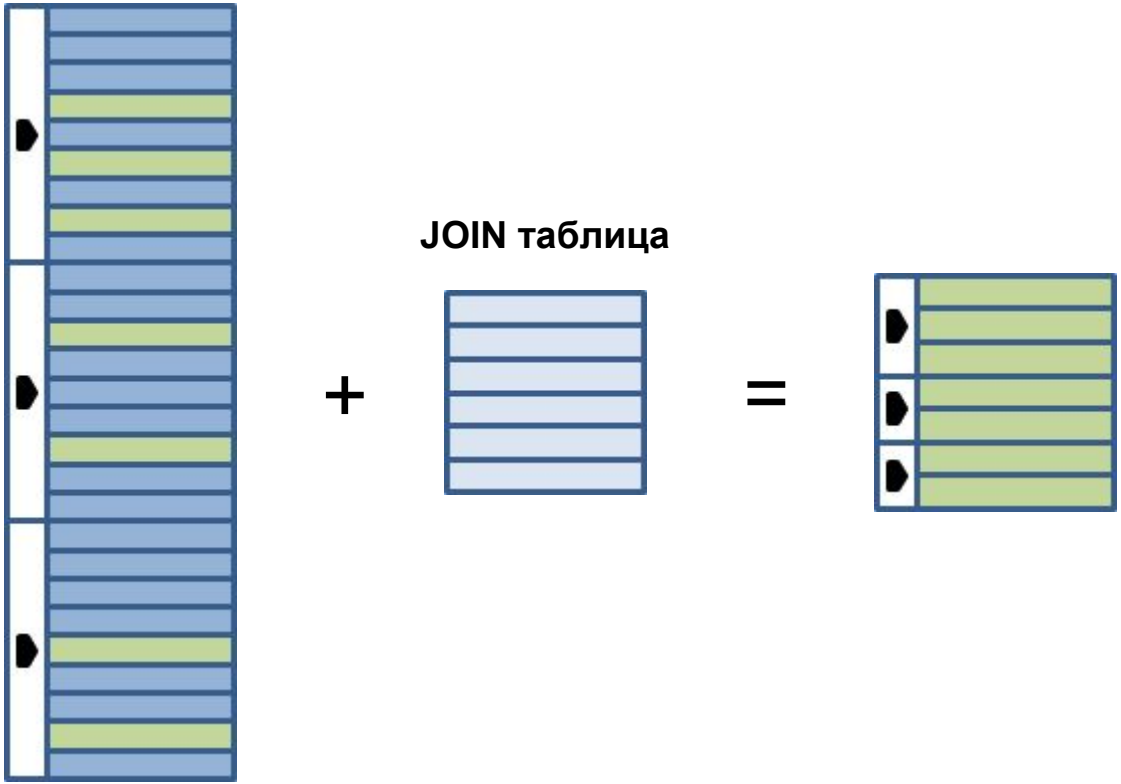
Пользователь загрузил ошибочные данные

Загружаем в CH с новым идентификатором, отмечаем удалённые в PSQL

Чистим БД время от времени

А если не нужна вся база?

Join таблицы
быстрое создание специализированной таблицы на лету



Итого:

- Генетика похожа на метрику
- Используйте связки из БД для гибкости
- ClickHouse действительно быстрая база (хочется индексов)
- Используйте таблицы OLD_ID -> NEW_ID для эмуляции удаления
- Эффект от загрузки блоками большой, используйте его!
- Join-таблицы фантастически быстрые (даже без индекса)
- Переносите математику в CH, у него много функций!
- ???