

Петабайт в YDB over HDD в процессингах Яндекс Метрики

Антон Барабанов
Яндекс Реклама

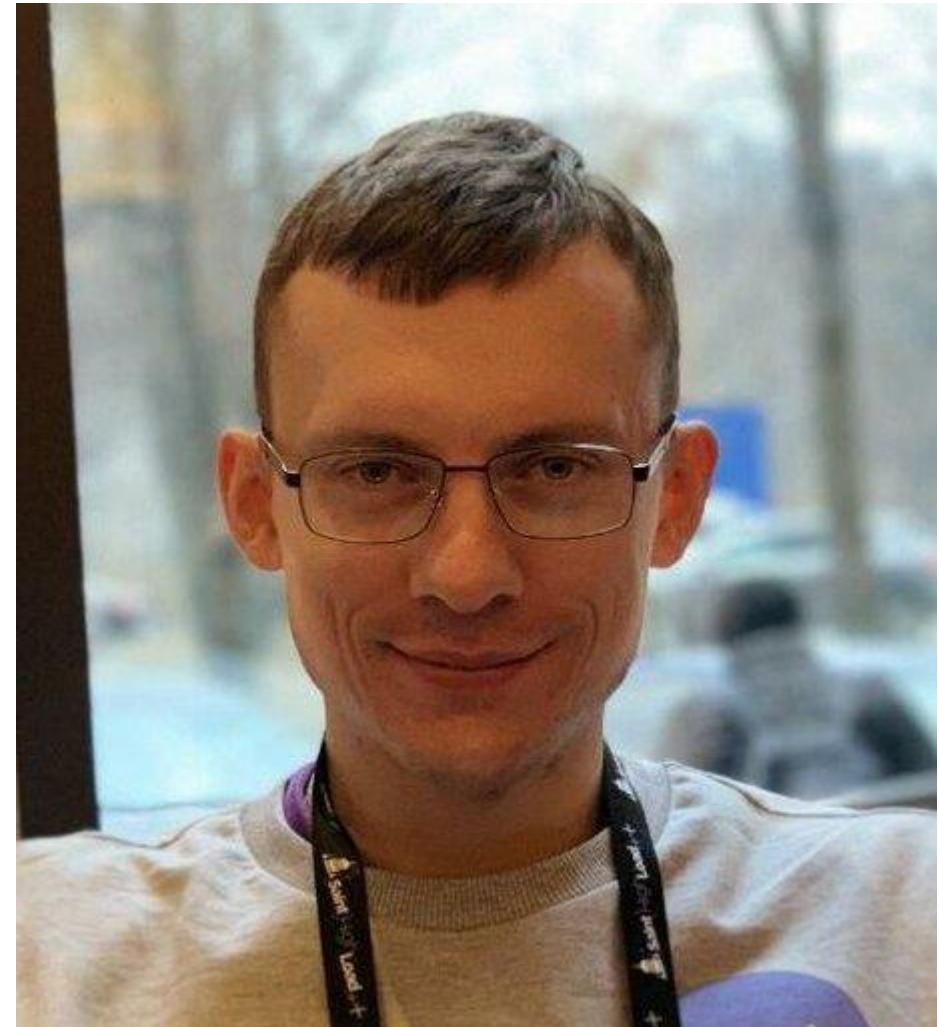


HighLoad ++



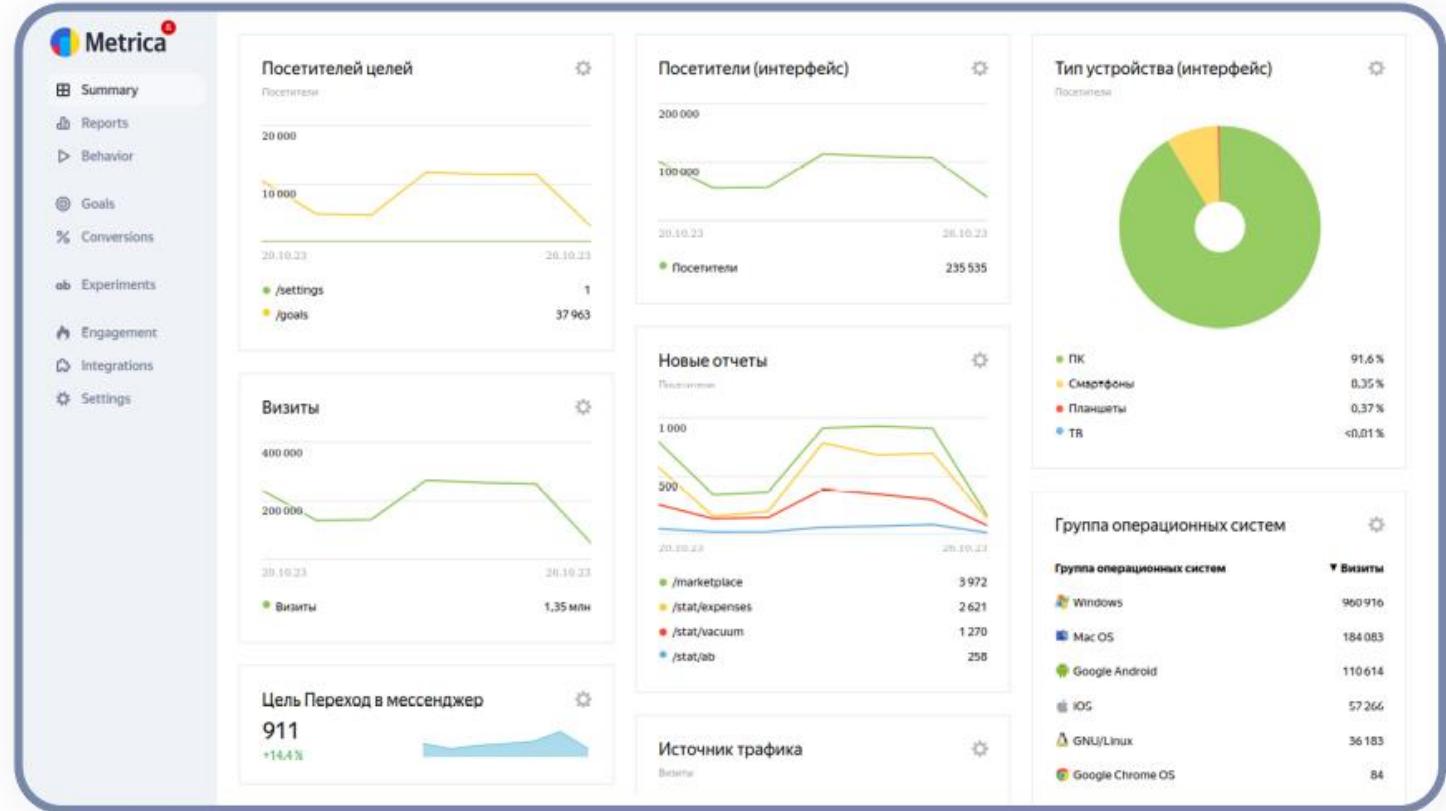
О себе

- Руковожу ядром Метрики
- 15 лет занимаюсь highload-системами
- Люблю все отказоустойчивое и расширяемое



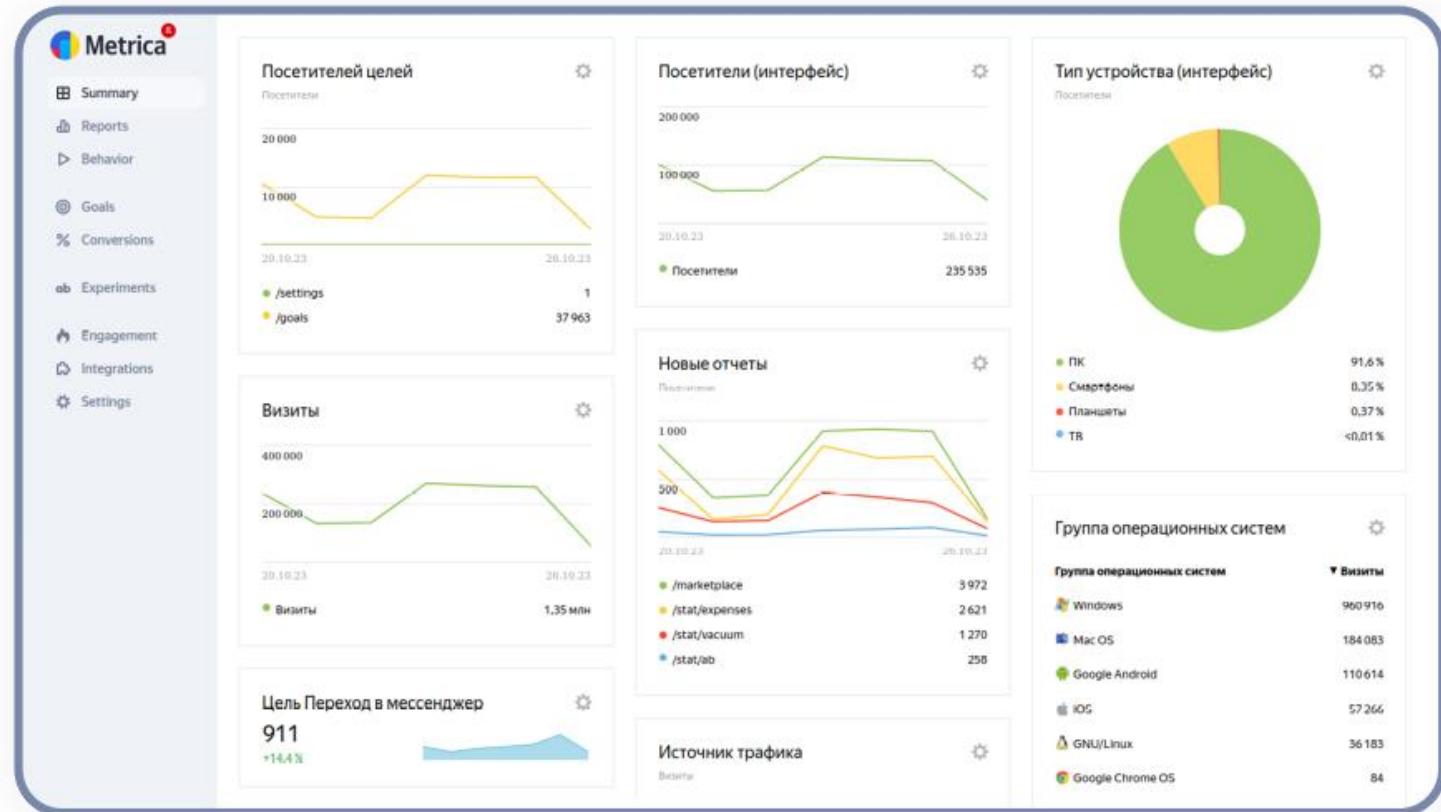
Метрика

- Самая популярная система аналитики в СНГ



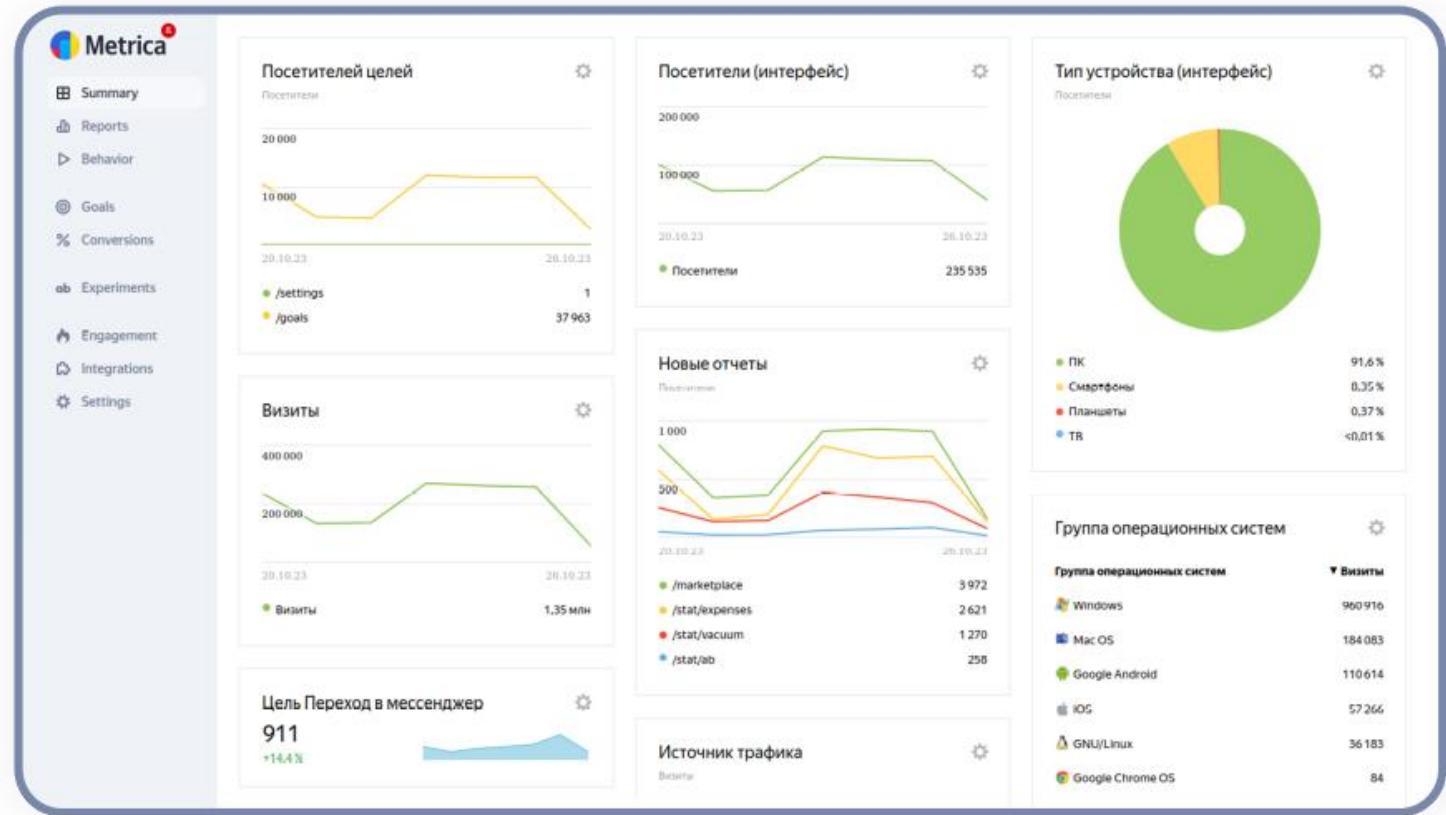
Метрика

- Самая популярная система аналитики в СНГ
- Третья по популярности в мире (w3techs.com 2019)



Метрика

- Самая популярная система аналитики в СНГ
- Третья по популярности в мире (w3techs.com 2019)
- Принимает от кода счетчика хиты, умеет записывать действия на сайте через вебвизор



Метрика

- Аккумулирует все сигналы в визиты

The screenshot shows the Yandex Metrica interface with a sidebar menu and a main session history area.

Sidebar Menu:

- Summary
- Reports
- Behavior
- Goals
- Conversions
- Experiments
- Engagement
- Integrations
- Settings

Main Area:

30 октября 2023

Визит в 09:53:57 | 6 | результаты поиска | 4 стр. | 25:48 | Подробнее

Сессия включает следующие события:

- Страница входа
- Просмотр
- Достигнута цель просмотр 2x страниц
- Просмотр
- Достигнута цель Просмотр 3x страниц
- Просмотр
- Достигнута цель Просматривал построенные
- Достигнута цель Просмотр 4x страниц
- Достигнута цель ЗАЯВКА (ВСЕ ФОРМЫ)
- Достигнута цель Заявка (все формы) новая

Метрика

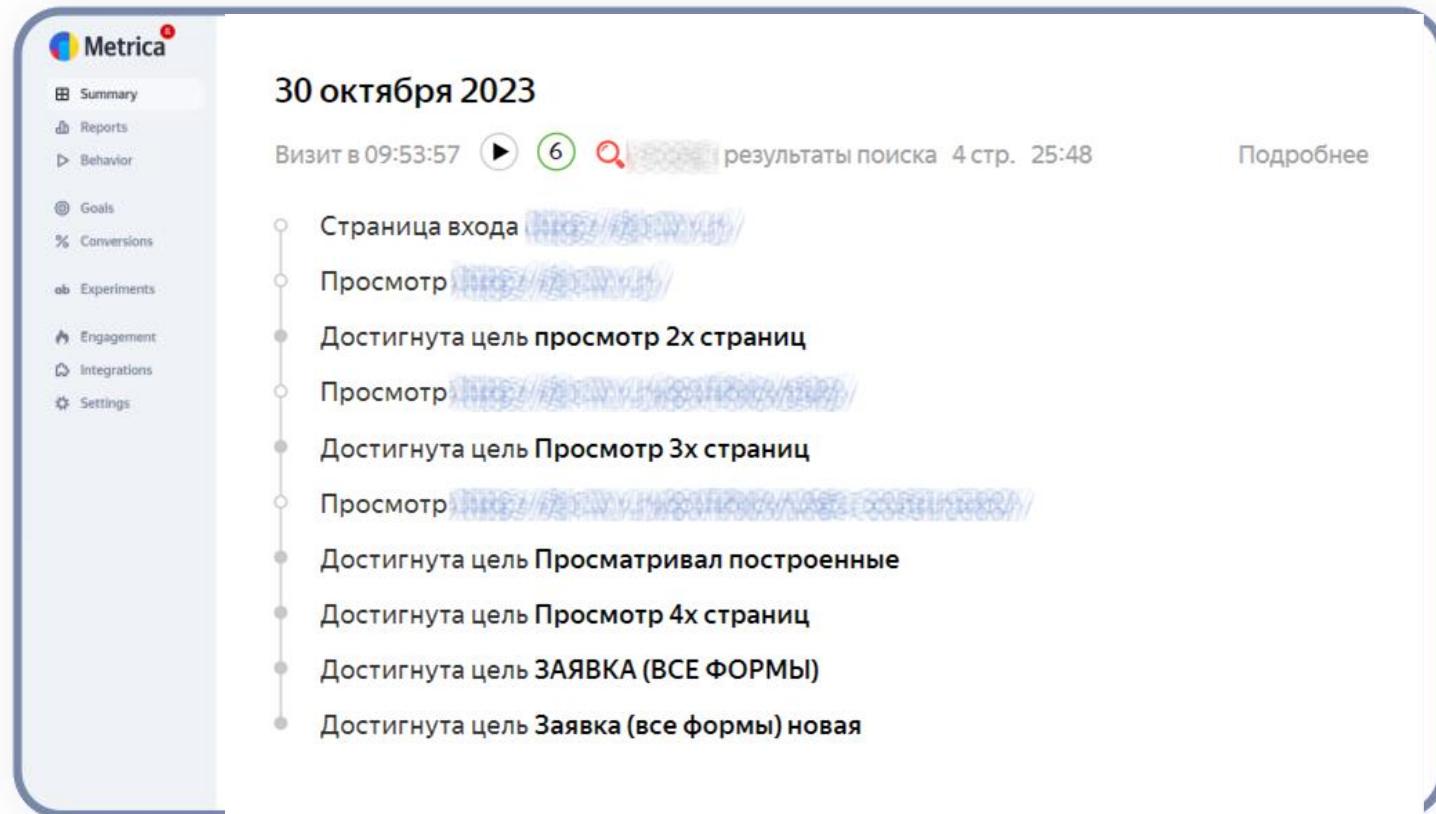
- Аккумулирует все сигналы в визиты
- Определяет источник трафика – поиск, реклама, прямой заход

The screenshot shows a visitor's session history from October 30, 2023, at 09:53:57. The session ID is 6. The visitor searched for 'результаты поиска' (search results) and viewed 4 pages in 25:48 minutes. The session details are as follows:

- Страница входа (Entry page)
- Просмотр (View)
- Достигнута цель просмотр 2x страниц (Goal: Viewed 2 pages)
- Просмотр (View)
- Достигнута цель Просмотр 3x страниц (Goal: Viewed 3 pages)
- Просмотр (View)
- Достигнута цель Просматривал построенные (Goal: Viewed built)
- Достигнута цель Просмотр 4x страниц (Goal: Viewed 4 pages)
- Достигнута цель ЗАЯВКА (ВСЕ ФОРМЫ) (Goal: Application (All Forms))
- Достигнута цель Заявка (все формы) новая (Goal: New Application (all forms))

Метрика

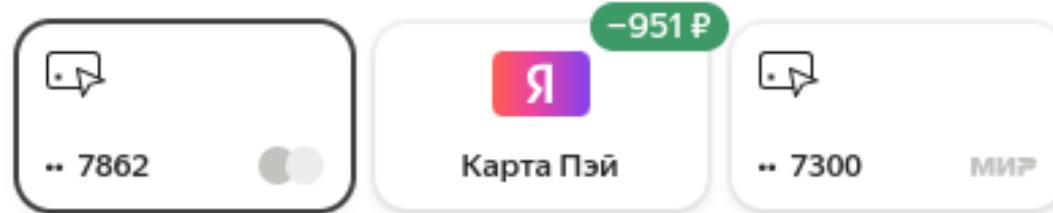
- Аккумулирует все сигналы в визиты
- Определяет источник трафика – поиск, реклама, прямой заход
- Строит продвинутые атрибуции на основании истории



Цели в метрике

- Они же конверсии

Способ оплаты

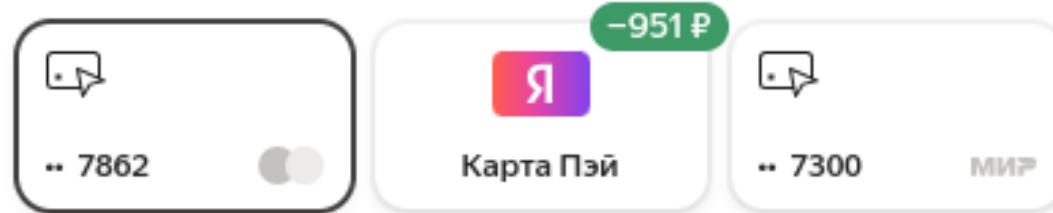
[Все способы](#)[Оплатить картой](#)

Безопасная оплата с Я Pay

Цели в метрике

- Они же конверсии
- Выделение ключевых действий из всего потока сигналов

Способ оплаты

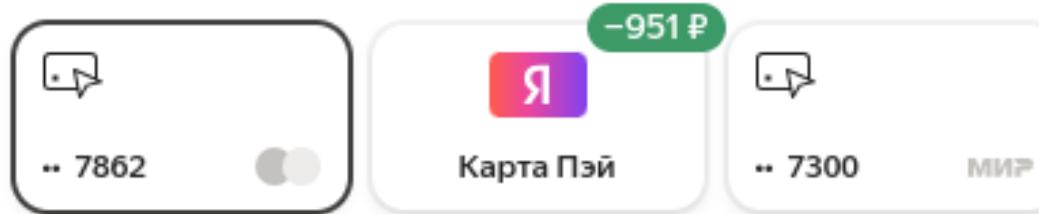
[Все способы](#)[Оплатить картой](#)

Безопасная оплата с Я Pay

Цели в метрике

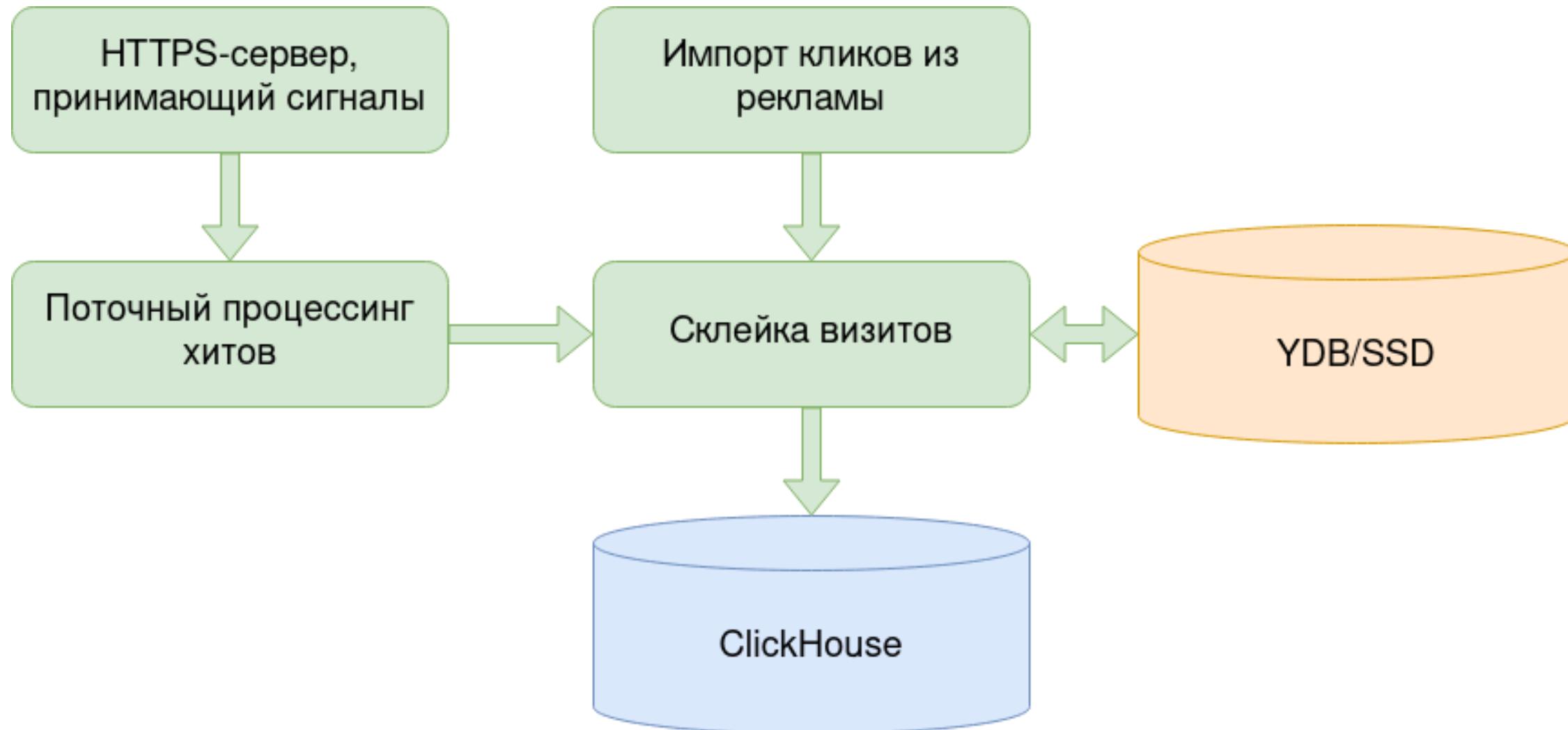
- Они же конверсии
- Выделение ключевых действий из всего потока сигналов
- Использование в рекламе

Способ оплаты

[Все способы](#)[Оплатить картой](#)

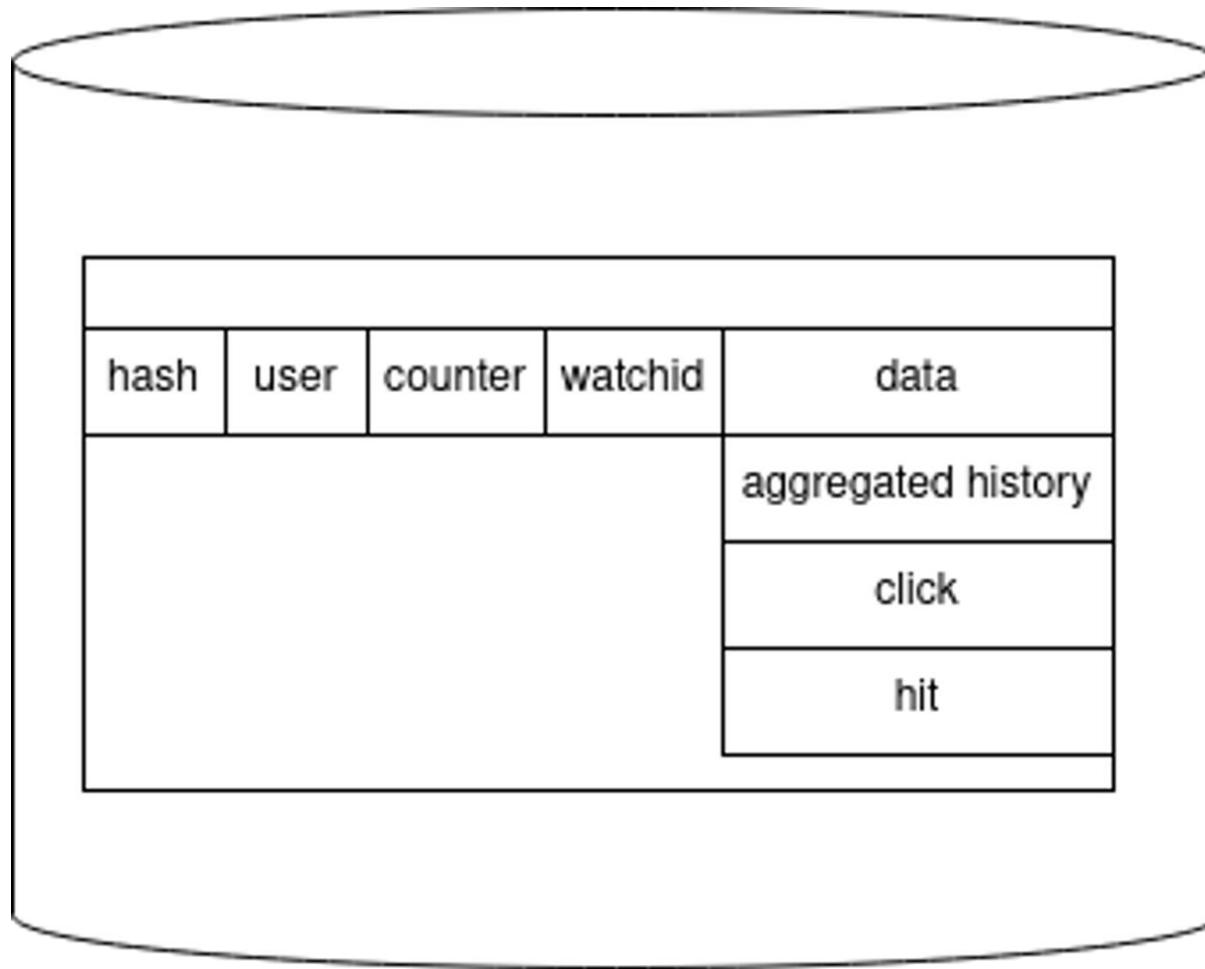
Безопасная оплата с Я Pay

Процессинг Метрики



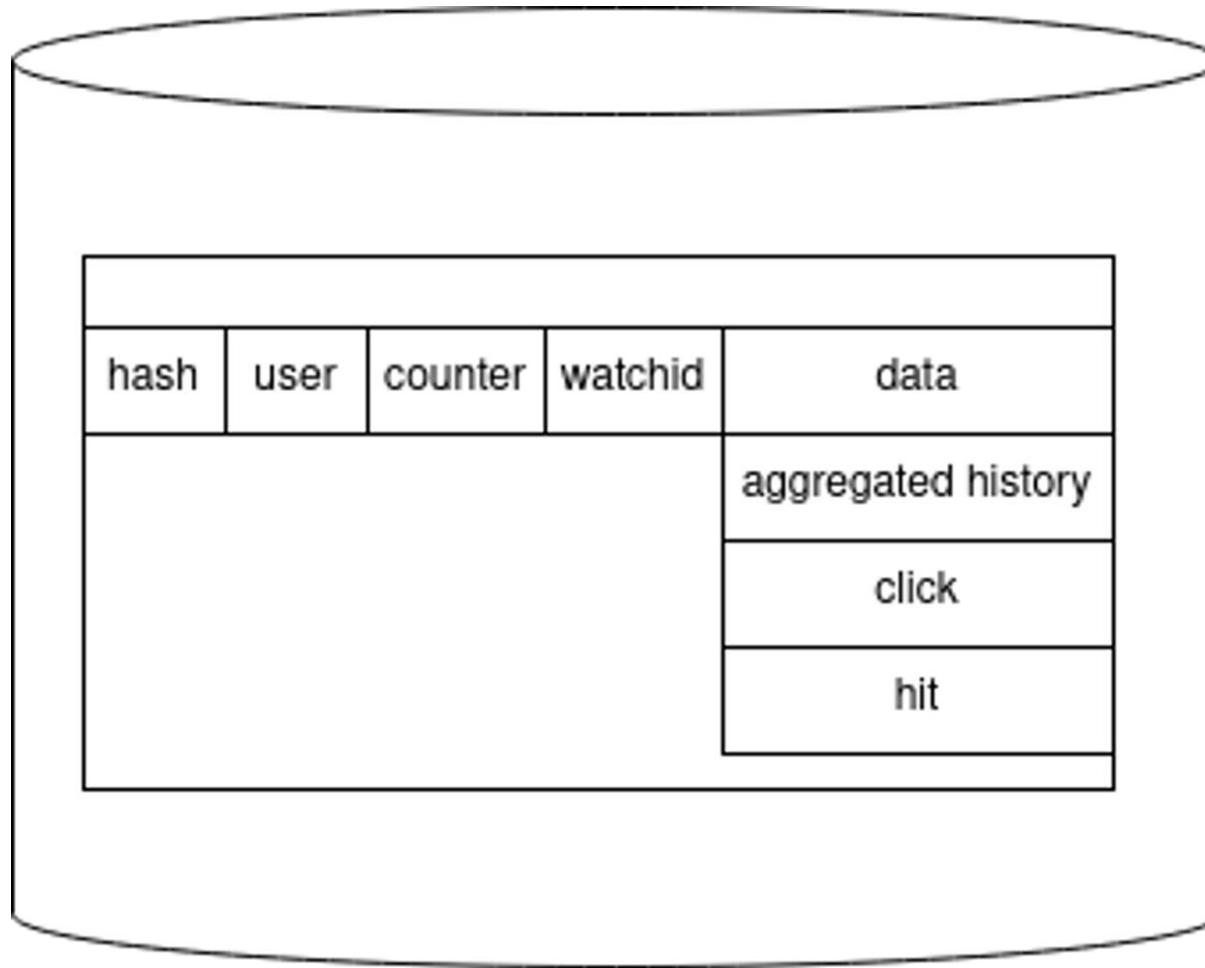
YDB в Метрике

- Дедупликация



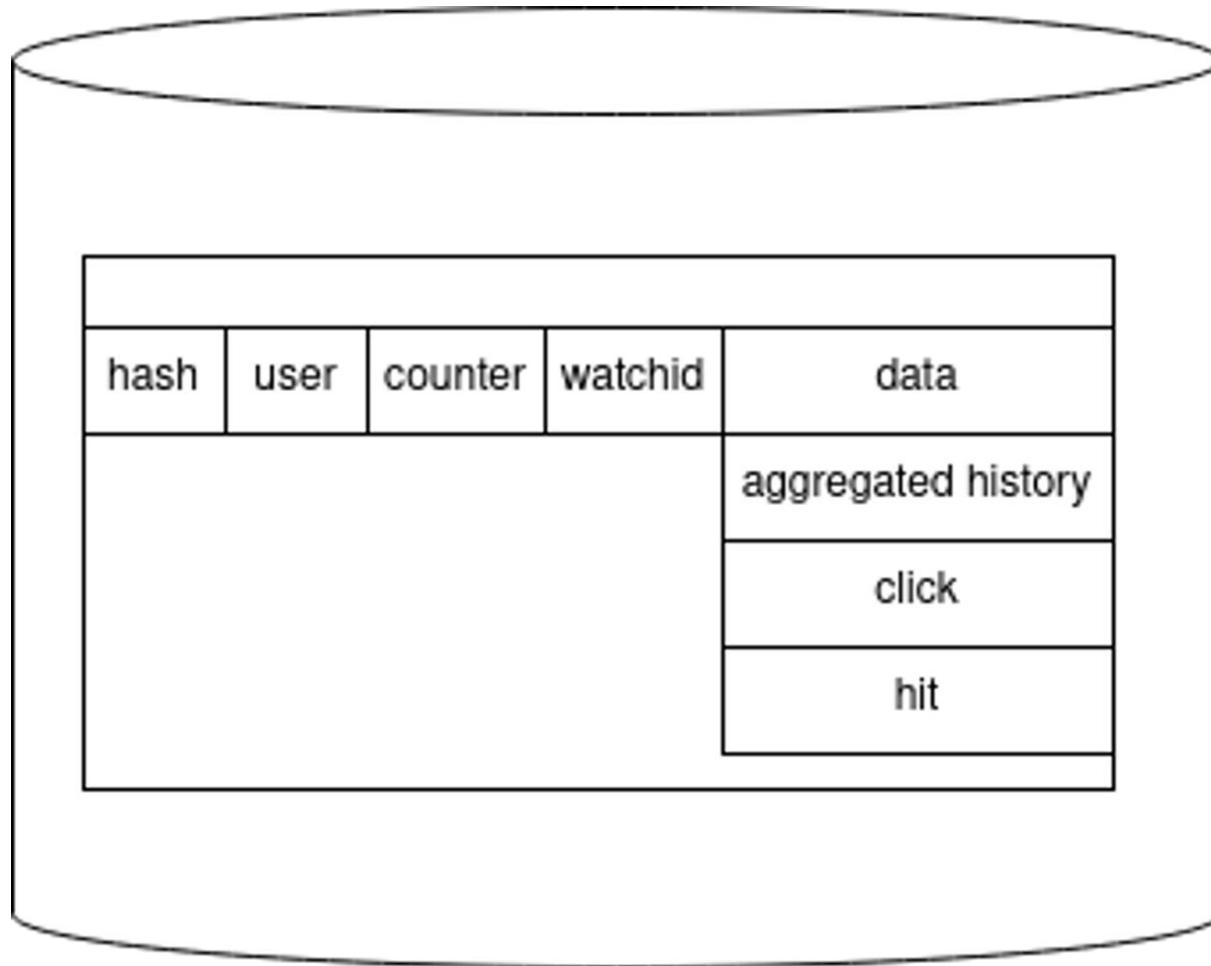
YDB в Метрике

- Дедупликация
- Сборка визитов



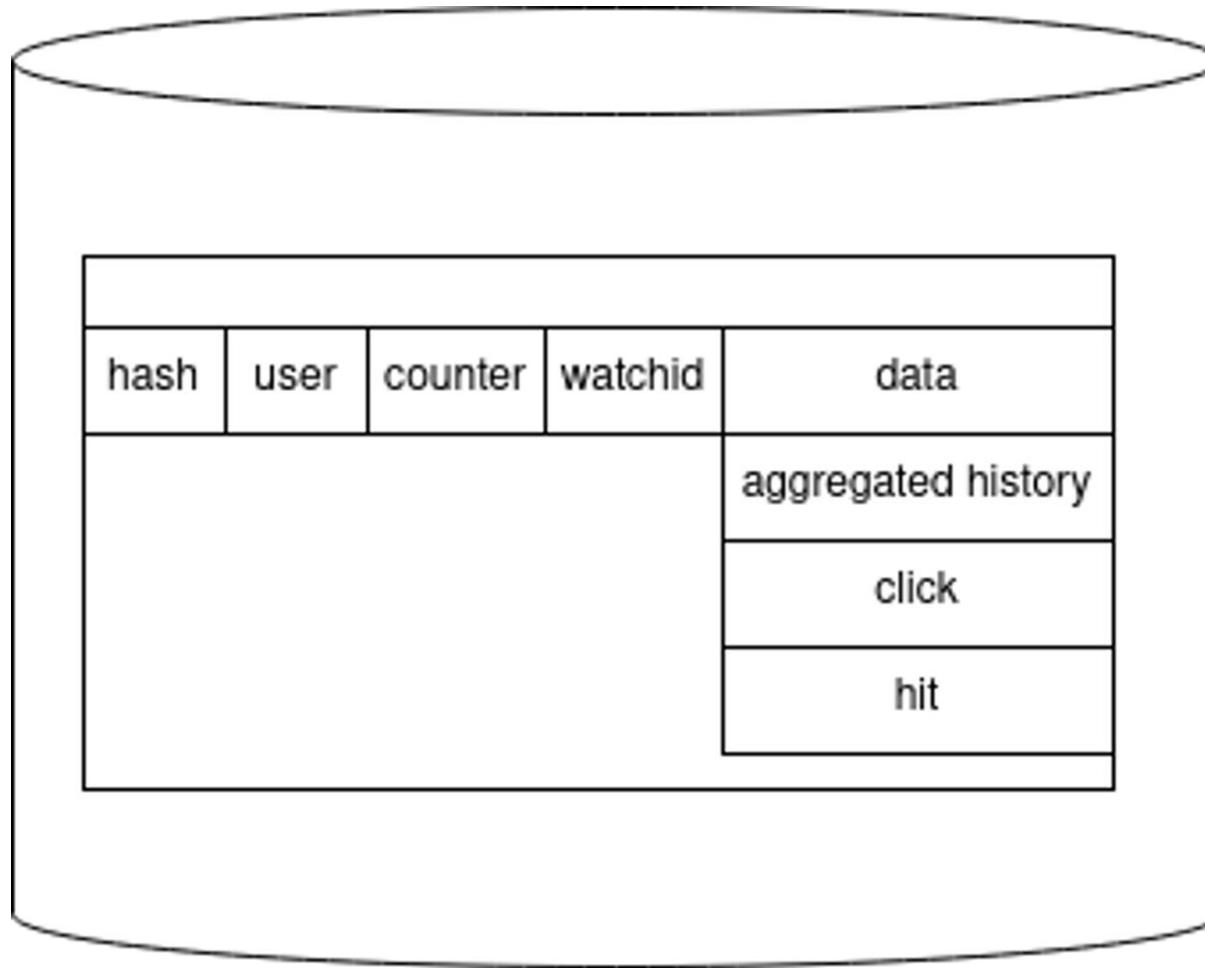
YDB в Метрике

- Дедупликация
- Сборка визитов
- TTL 24 часа



YDB в Метрике

- Дедупликация
- Сборка визитов
- TTL 24 часа
- 100 терабайт



Как мы работаем с YDB

- В корне базы лежат папки



Как мы работаем с YDB

- В корне базы лежат папки
- В папках – таблицы



Как мы работаем с YDB

- В корне базы лежат папки
- В папках – таблицы
- Таблицы разбиты на партиции



Как мы работаем с YDB

- В корне базы лежат папки
- В папках – таблицы
- Таблицы разбиты на партиции
- Распределенность базы достигается за счет партиций



Как мы работаем с YDB

- Партиции разбиваются по первичному ключу

Как мы работаем с YDB

- Партиции разбиваются по первичному ключу
- Запрос в рамках одной партиции - это локальная транзакция

Как мы работаем с YDB

- Партиции разбиваются по первичному ключу
- Запрос в рамках одной партиции - это локальная транзакция
- Если локализации нет - запускается распределенная транзакция



Как мы работаем с YDB

- Партиции разбиваются по первичному ключу
- Запрос в рамках одной партиции - это локальная транзакция
- Если локализации нет – запускается распределенная транзакция
- BulkUpsert позволяет вставлять данные, минуя транзакции вовсе



Как мы работаем с YDB

- При создании таблицы можно указать стартовое число партиций, минимальное и максимальное

Как мы работаем с YDB

- При создании таблицы можно указать стартовое число партиций, минимальное и максимальное
- Также можно указать максимальный размер партиции в байтах

Как мы работаем с YDB

- При создании таблицы можно указать стартовое число партиций, минимальное и максимальное
- Также можно указать максимальный размер партиции в байтах
- Есть 2 вида автосплита: по нагрузке и по размеру

Как мы работаем с YDB

- При создании таблицы можно указать стартовое число партиций, минимальное и максимальное
- Также можно указать максимальный размер партиции в байтах
- Есть 2 вида автосплита: по нагрузке и по размеру
- Для работы автосплита важно, чтобы были указаны min и max, и чтобы min был строго меньше max



Конверсии из реального мира



Конверсии из реального мира

- Большая глубина поиска – нужно хранить все визиты несколько десятков дней

Конверсии из реального мира

- Большая глубина поиска – нужно хранить все визиты несколько десятков дней
- Нет строгих требований к realtime – проклейка за 15-20 минут вполне подойдет

Конверсии из реального мира

- Большая глубина поиска – нужно хранить все визиты несколько десятков дней
- Нет строгих требований к realtime – проклейка за 15-20 минут вполне подойдет
- Есть определенные требования к realtime – клеить 15-20 часов уже плохо

Требования к БД для хранения

- Около 40 миллиардов строк в сутки размером около 2 килобайт

Требования к БД для хранения

- Около 40 миллиардов строк в сутки размером около 2 килобайт
- Нужно хранить на HDD, чтобы не простоявали сотни терабайт дорогих SSD

Требования к БД для хранения

- Около 40 миллиардов строк в сутки размером около 2 килобайт
- Нужно хранить на HDD, чтобы не простоявали сотни терабайт дорогих SSD
- Нужно уметь доставать данные по ключу за разумное время

Требования к БД для хранения

- Около 40 миллиардов строк в сутки размером около 2 килобайт
- Нужно хранить на HDD, чтобы не простоявали сотни терабайт дорогих SSD
- Нужно уметь доставать данные по ключу за разумное время
- Высокие стабильность и предсказуемость

Варианты базы

ClickHouse

- Хорошо умеет работать на HDD

Варианты базы

ClickHouse

- Хорошо умеет работать на HDD
- Большое community, хороший личный контакт

Варианты базы

ClickHouse

- Хорошо умеет работать на HDD
- Большое community, хороший личный контакт
- Большой опыт эксплуатации в Метрике

Варианты базы

ClickHouse

- Хорошо умеет работать на HDD
- Большое community, хороший личный контакт
- Большой опыт эксплуатации в Метрике

YDB

- Отлично работает с key-value-запросами

Варианты базы

ClickHouse

- Хорошо умеет работать на HDD
- Большое community, хороший личный контакт
- Большой опыт эксплуатации в Метрике

YDB

- Отлично работает с key-value-запросами
- Активно развивается, хороший личный контакт

Варианты базы

ClickHouse

- Хорошо умеет работать на HDD
- Большое community, хороший личный контакт
- Большой опыт эксплуатации в Метрике

YDB

- Отлично работает с key-value-запросами
- Активно развивается, хороший личный контакт
- Большой опыт эксплуатации в Метрике

Варианты базы

ClickHouse

✗ Плохо работает как key-value

Варианты базы

ClickHouse

- ✗ Плохо работает как key-value
- ✗ Плохо работает с синхронной репликацией

Варианты базы

ClickHouse

- ✗ Плохо работает как key-value
- ✗ Плохо работает с синхронной репликацией

YDB

- ✗ Неизвестно, как работает на HDD

Первая версия схемы

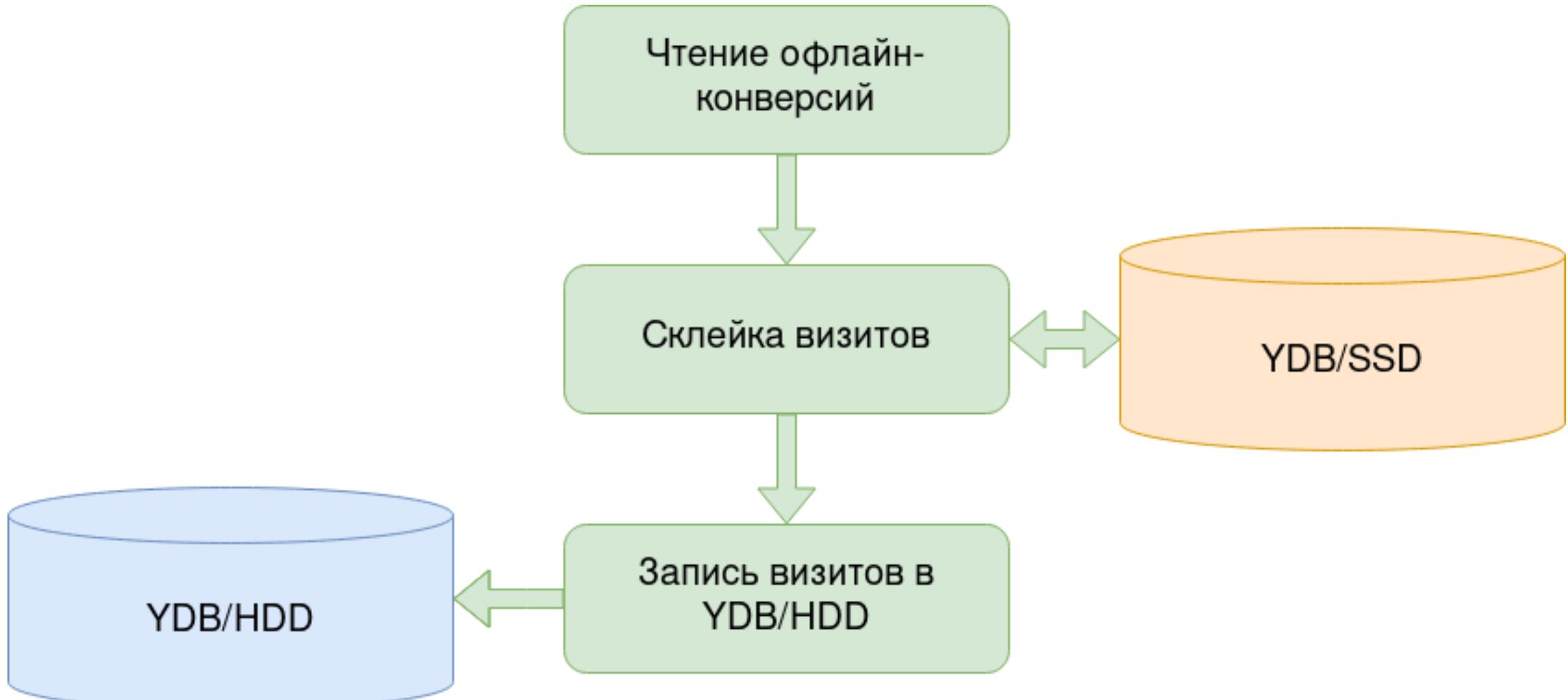


Схема данных

ID	Version	Data
1	1	qwerty

- В каждой строке есть первичный ключ – ID, версия и данные

Схема данных

ID	Version	Data
1	1	qwerty

- В каждой строке есть первичный ключ – ID, версия и данные
- Запись и чтение идет итерациями

Схема данных

ID	Version	Data
1	1	qwerty
1	2	qwerty

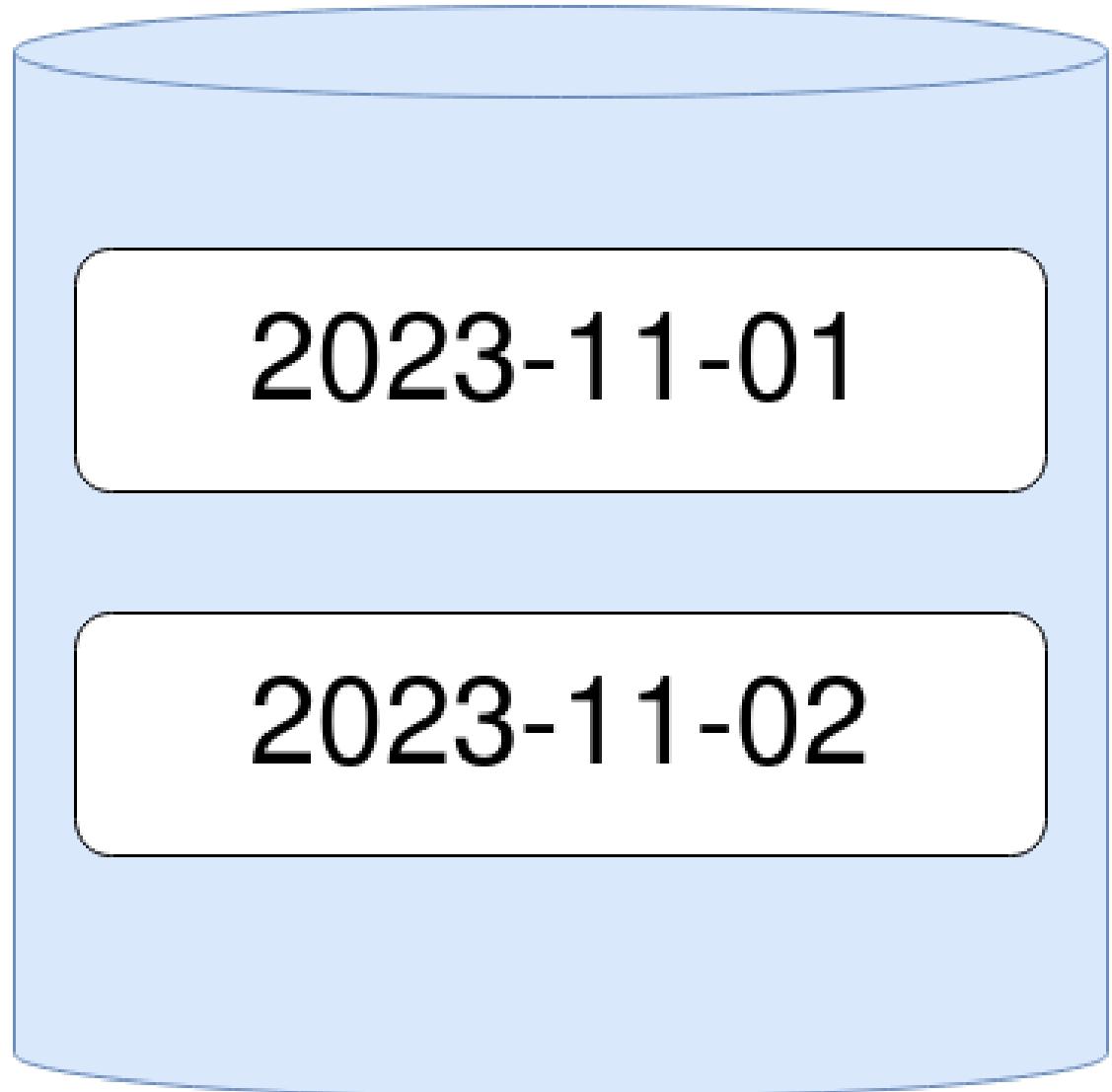
- Игнорируем все строки с версией \geq номера итерации

Схема данных

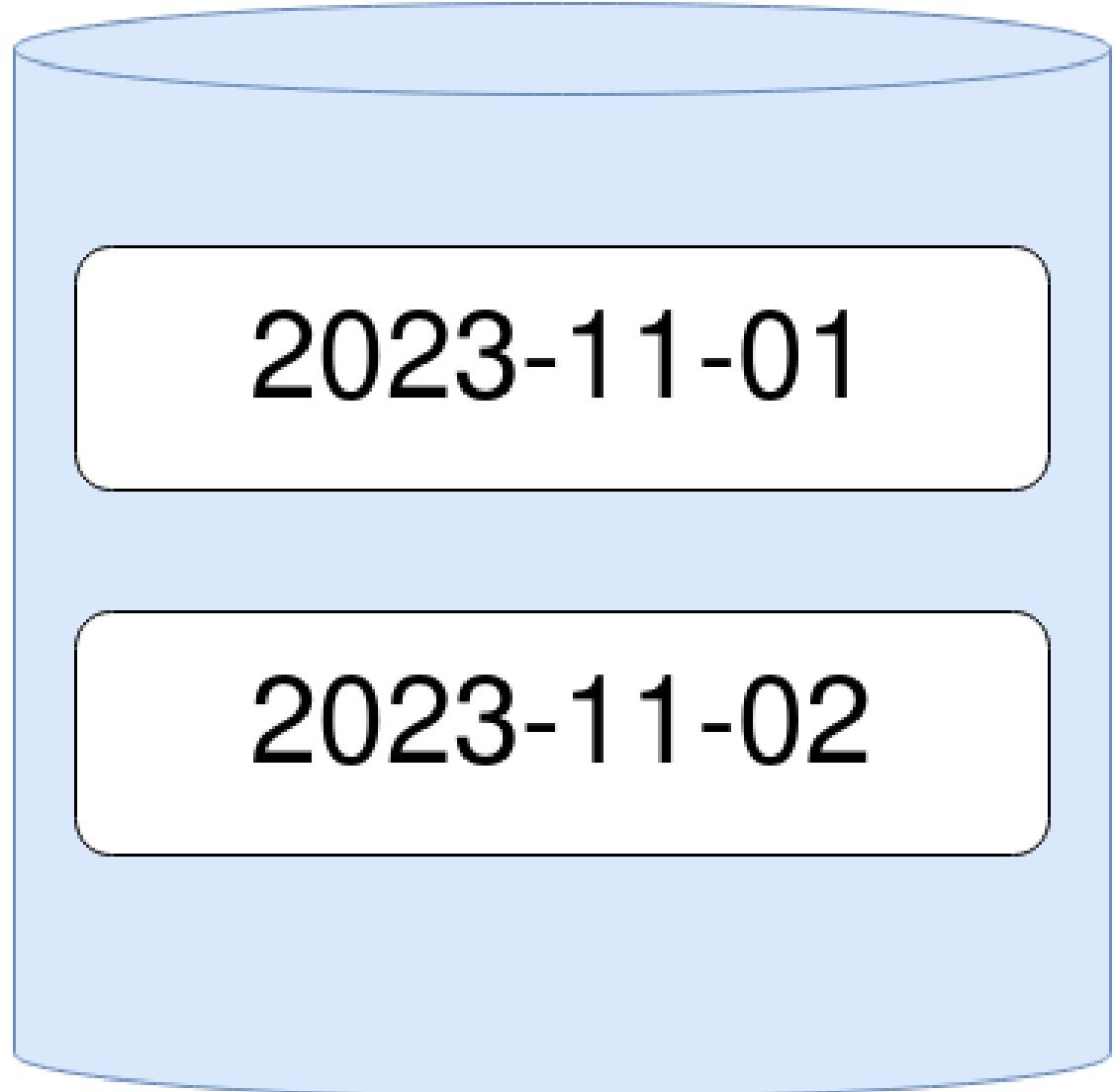
ID	Version	Data
1	1	qwerty
1	2	qwerty

- Игнорируем все строки с версией \geq номера итерации
- Удаляем старую версию после того, как проведем транзакцию в процессинге

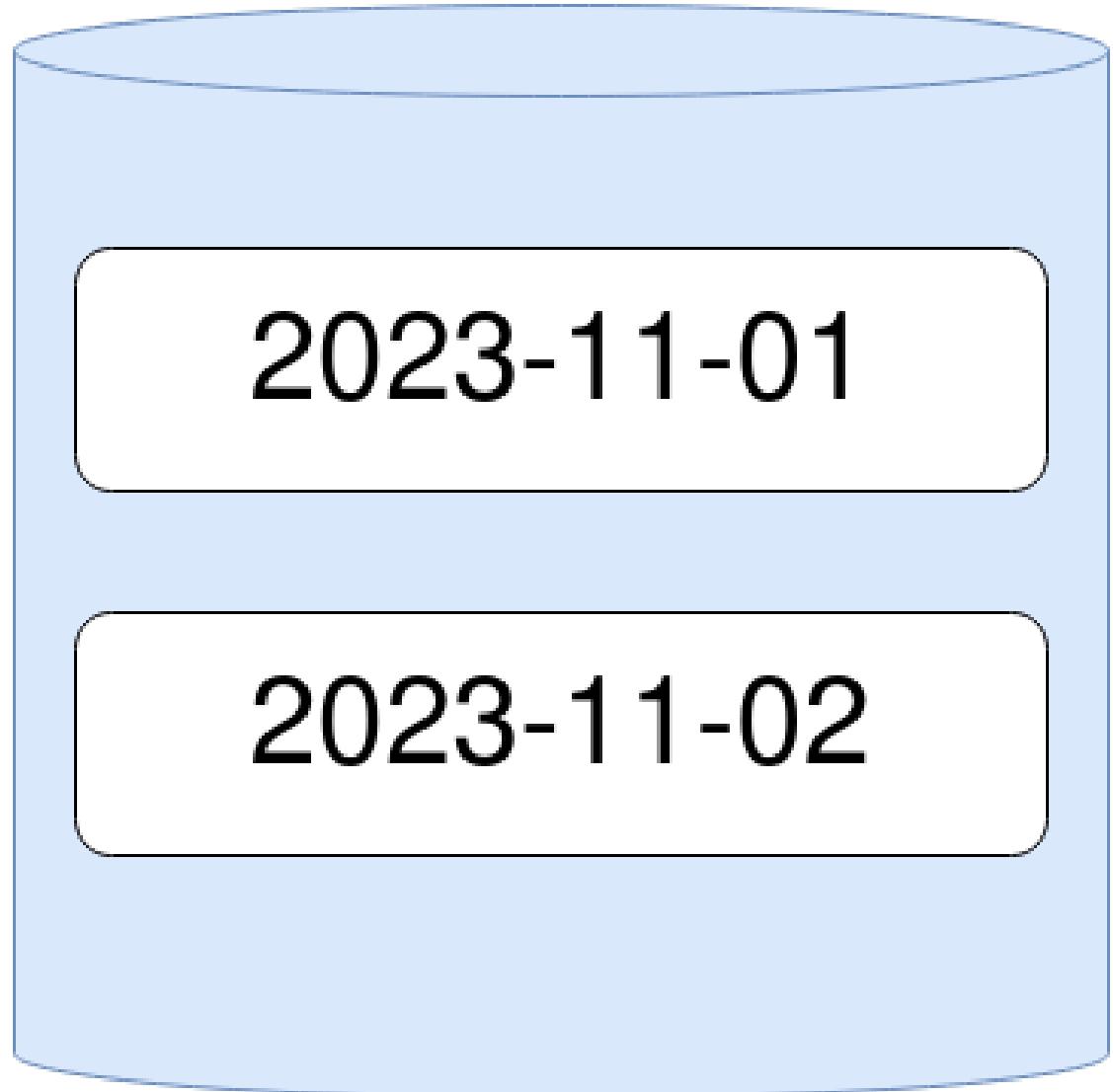
- Встроенный TTL делает регулярное полное сканирование таблицы



- Встроенный TTL делает регулярное полное сканирование таблицы
- Для размерностей в сотни терабайт не предназначен



- Встроенный TTL делает регулярное полное сканирование таблицы
- Для размерностей в сотни терабайт не предназначен
- Решение: хранить данные по дням.



Попытка включения 1

Попытка включения 1



Попытка включения 1

- Все легло

Попытка включения 1

- Все легло
- В YDB есть интервальная зависимость размера хранилища от требований к RAM

Попытка включения 1

- Все легко
- В YDB есть интервальная зависимость размера хранилища от требований к RAM
- В YDB достаточно активно применяется кеширование

Попытка включения 1

- Все легко
- В YDB есть интервальная зависимость размера хранилища от требований к RAM
- В YDB достаточно активно применяется кеширование
- И в нашей инсталляции есть жесткая привязка RAM к CPU

Попытка включения 1

- Все легко
- В YDB есть интервальная зависимость размера хранилища от требований к RAM
- В YDB достаточно активно применяется кеширование
- И в нашей инсталляции есть жесткая привязка RAM к CPU
- Из-за этого пришлось доливать вычислительные ресурсы

Попытка включения 1

- Все легко
- В YDB есть интервальная зависимость размера хранилища от требований к RAM
- В YDB достаточно активно применяется кеширование
- И в нашей инсталляции есть жесткая привязка RAM к CPU
- Из-за этого пришлось доливать вычислительные ресурсы
- Результат – текущая утилизация CPU около 8%

Попытка включения 2

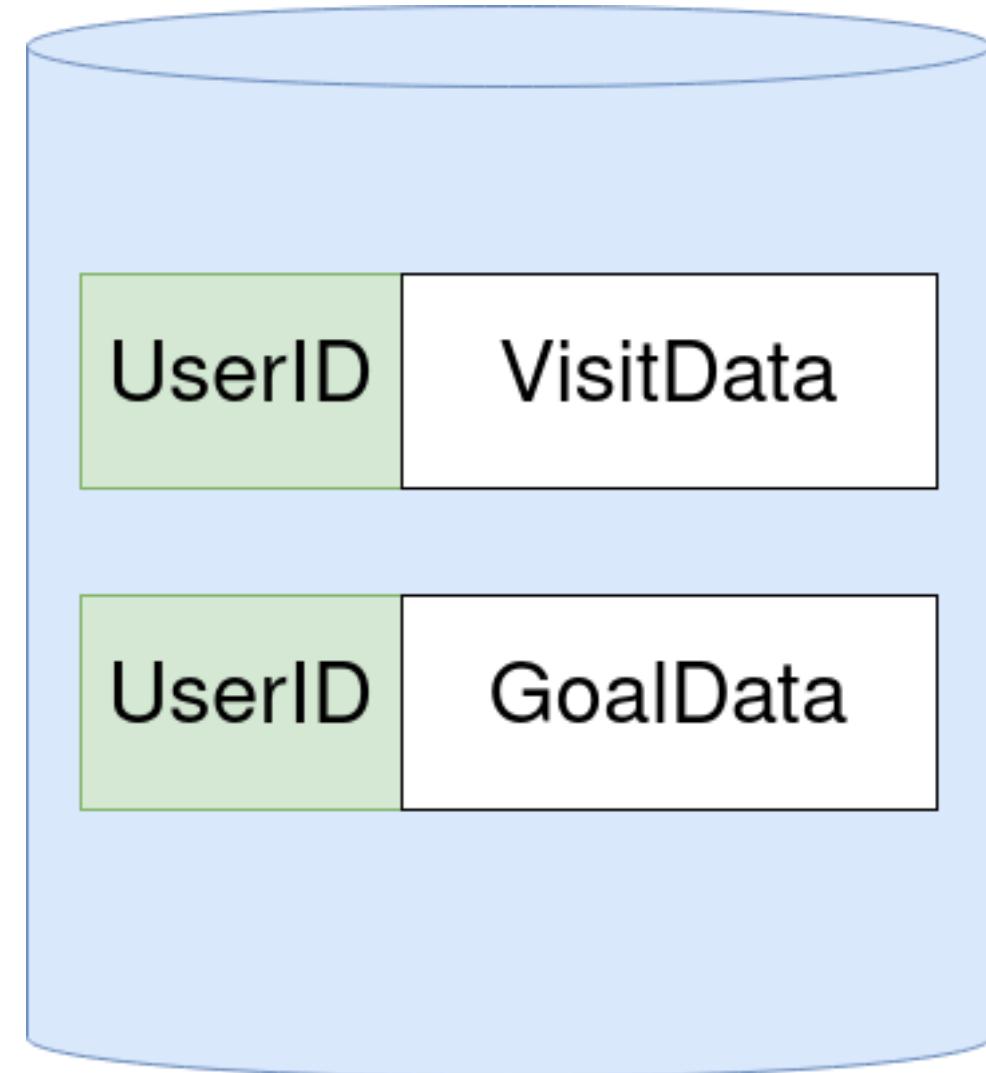
Попытка включения 2

63



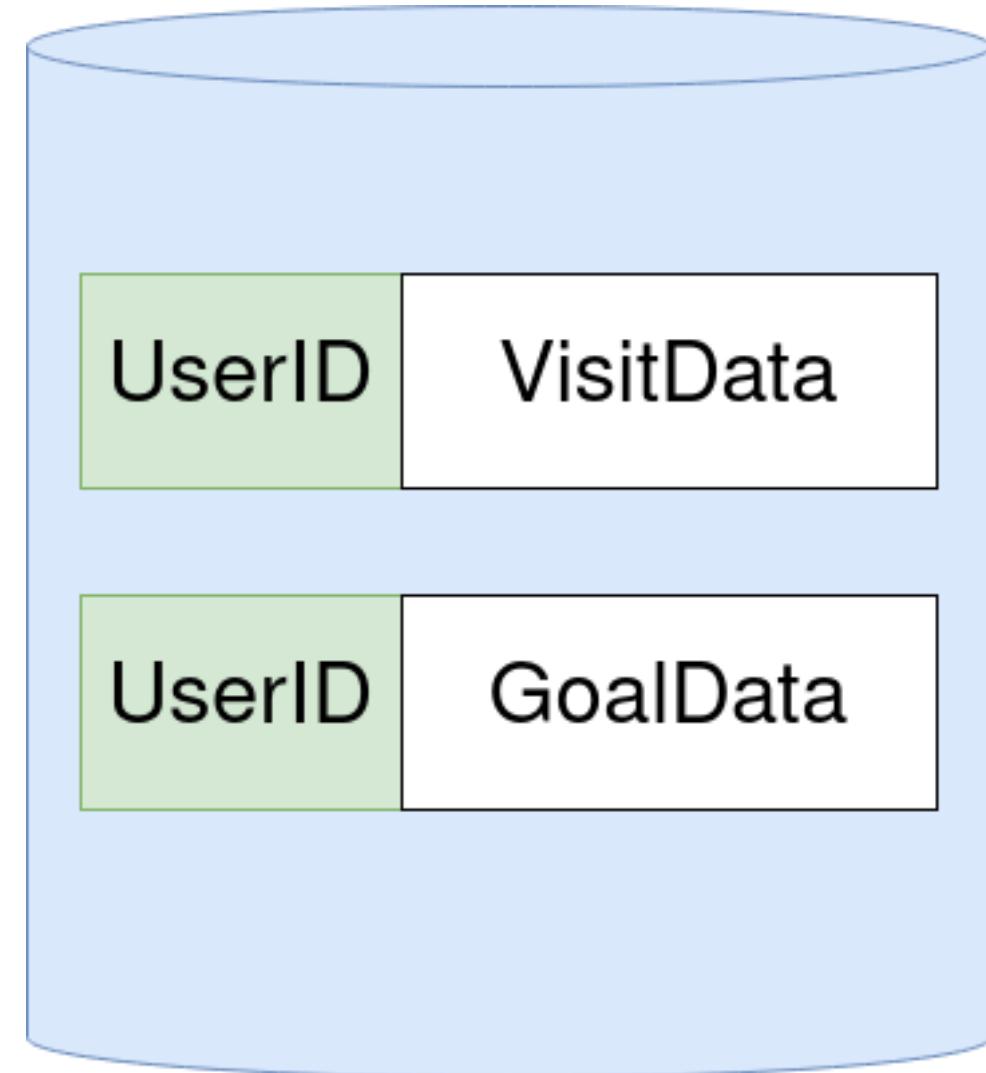
Попытка включения 2

- Все легло



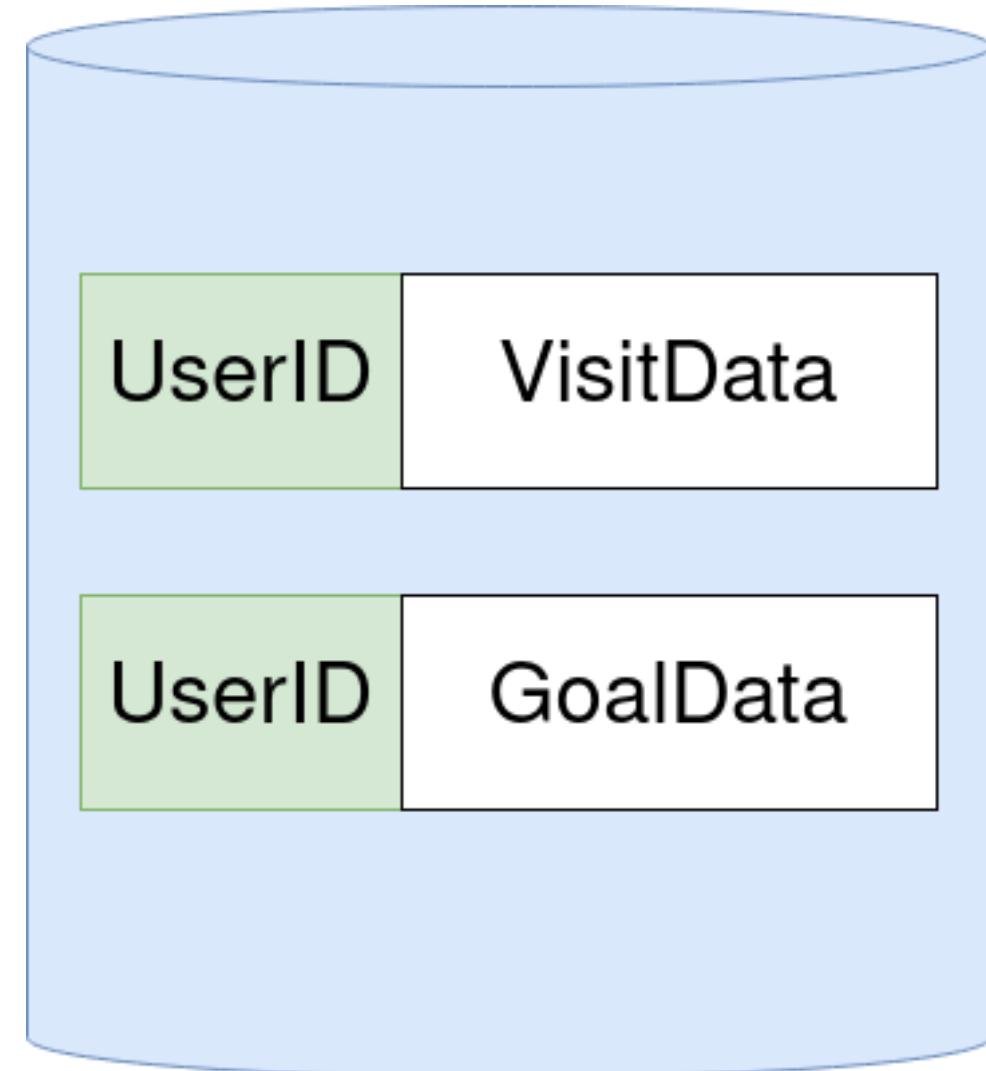
Попытка включения 2

- Все легко
- Диски HDD не справляются с большой нагрузкой на чтение



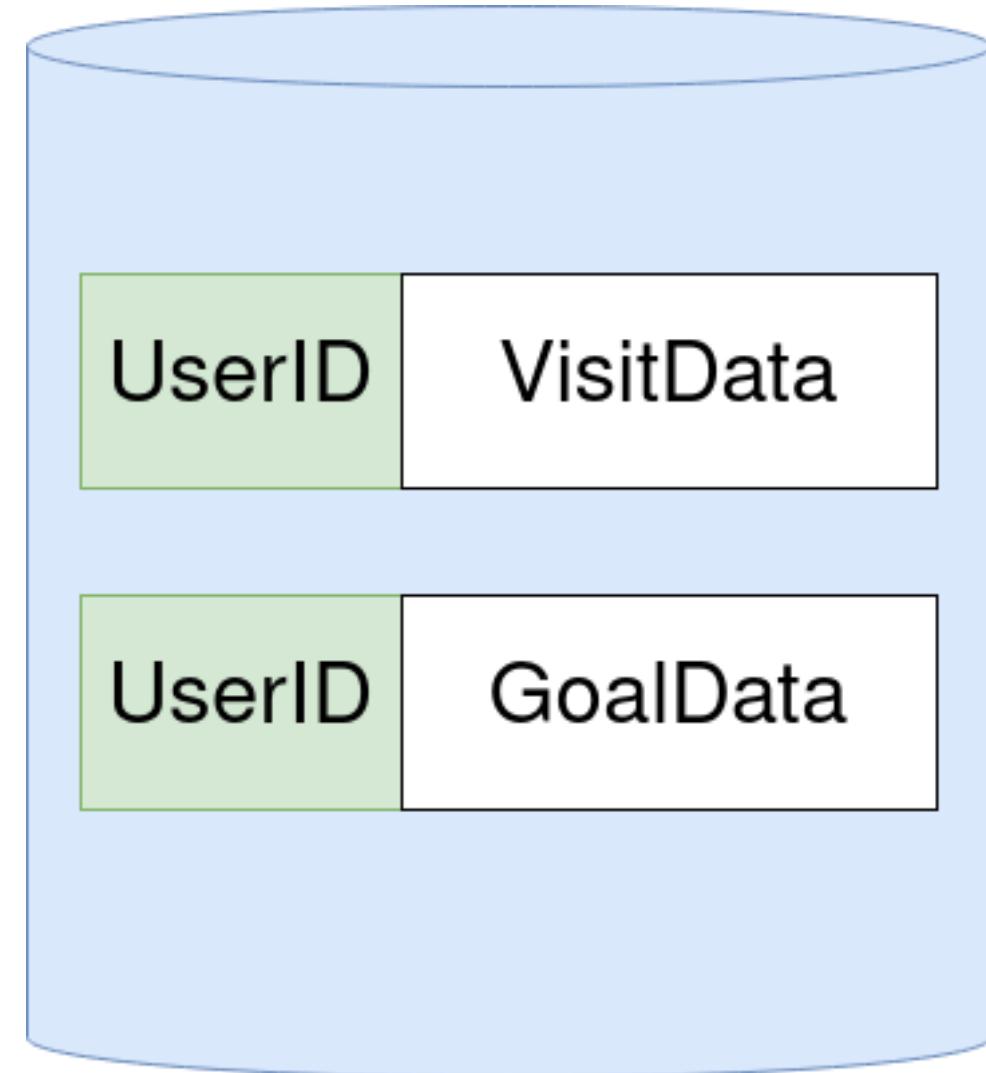
Попытка включения 2

- Все легко
- Диски HDD не справляются с большой нагрузкой на чтение
- YDB вычитывает блок и только потом понимает, что там нет данных



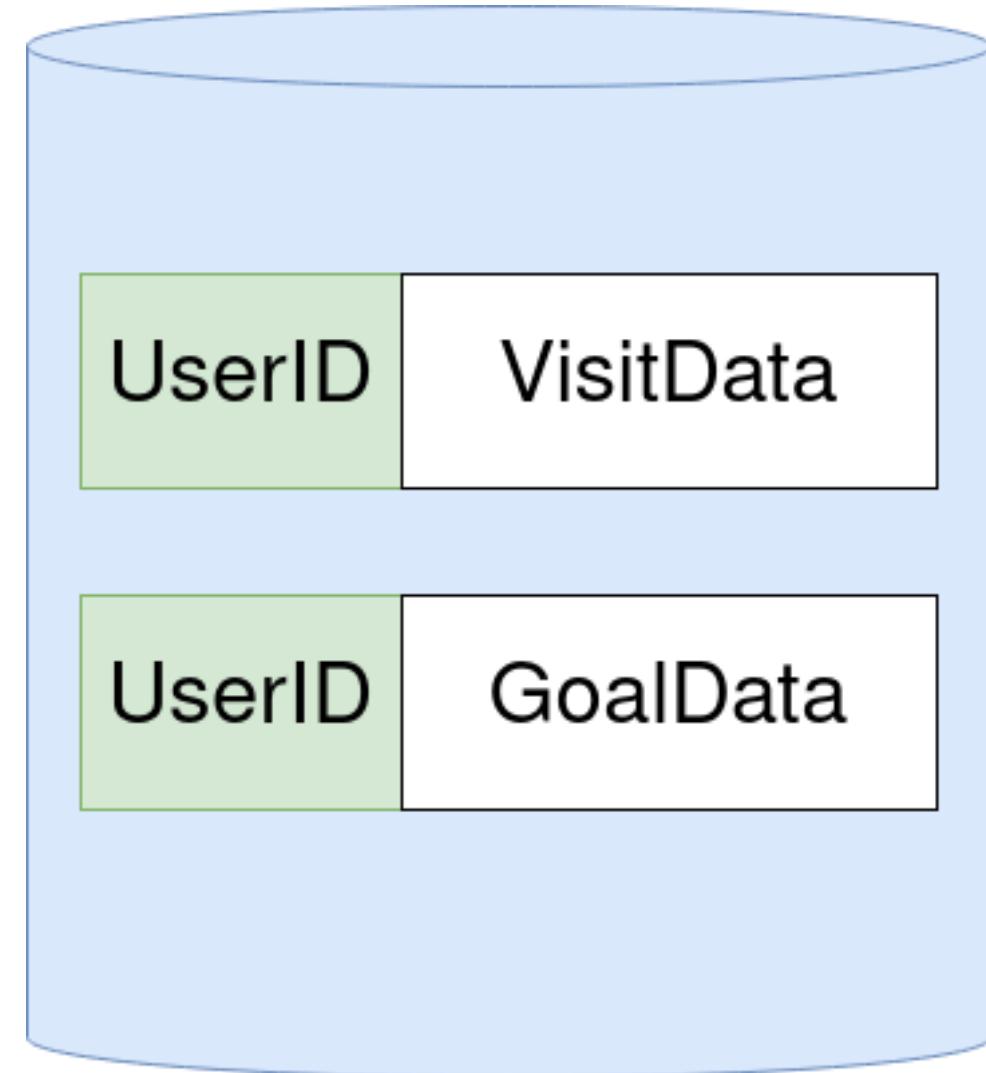
Попытка включения 2

- Встроенный bloom-фильтр не поможет, он работает по полному первичному ключу



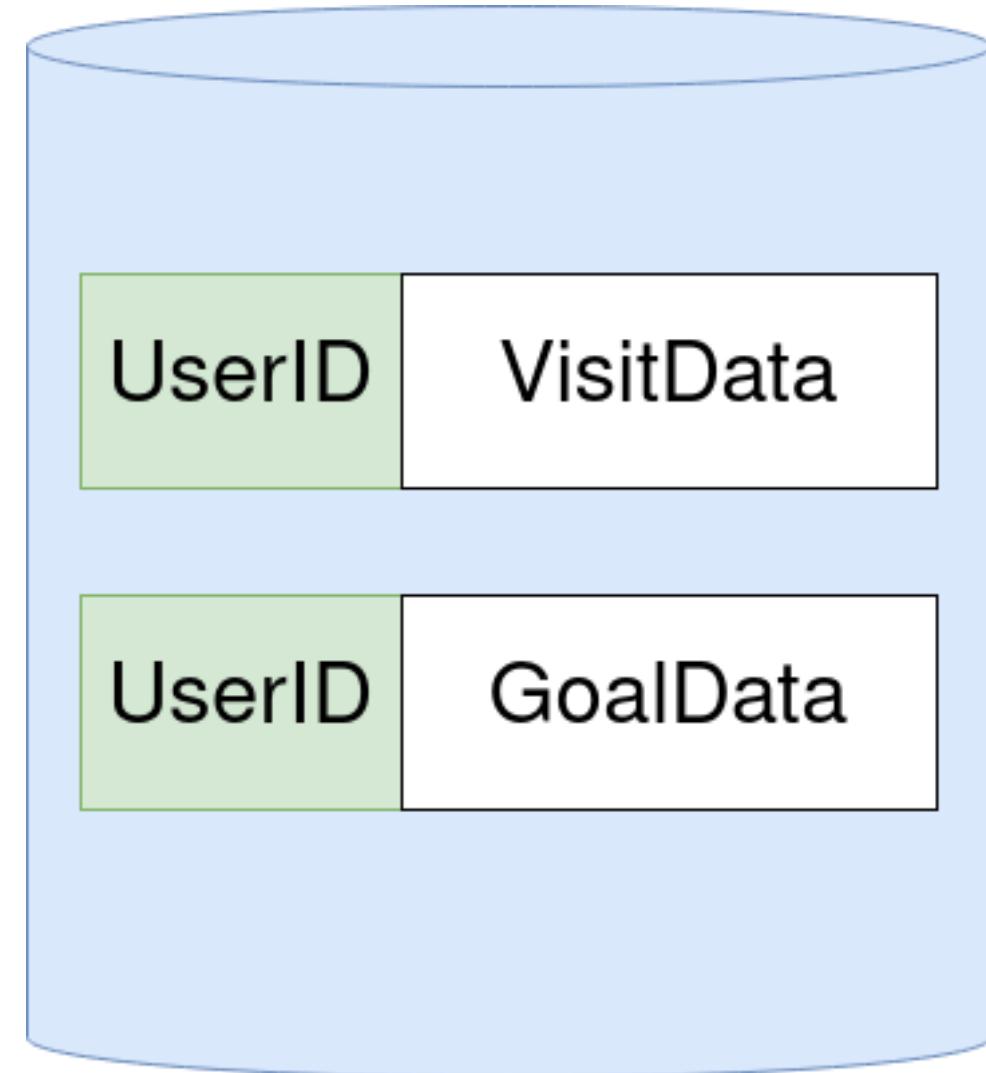
Попытка включения 2

- Встроенный bloom-фильтр не поможет, он работает поному первичному ключу
- Положим ключ на SSD, а данные оставим на HDD



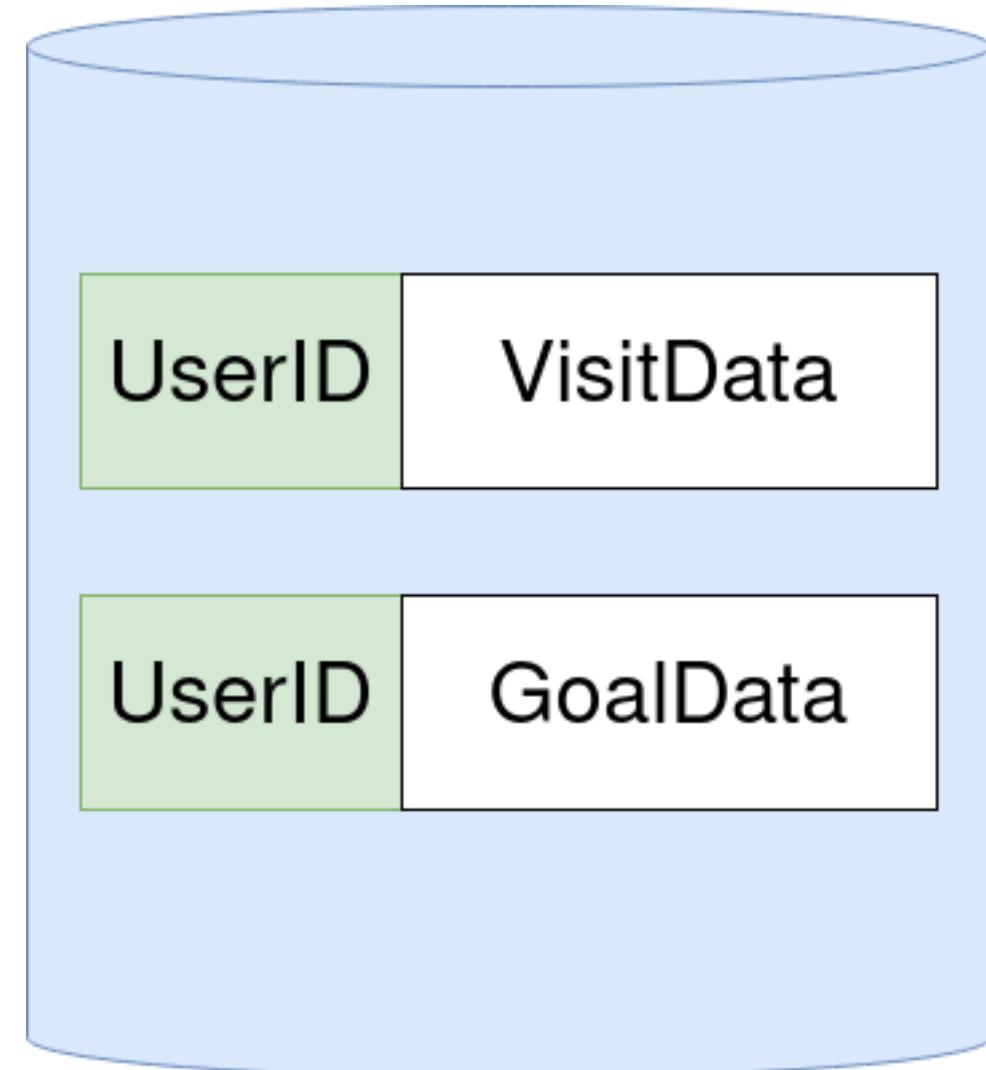
Попытка включения 2

- Встроенный bloom-фильтр не поможет, он работает поному первичному ключу
- Положим ключ на SSD, а данные оставим на HDD
- Первой частью запроса с SSD полный ключ



Попытка включения 2

- Встроенный bloom-фильтр не поможет, он работает поному первичному ключу
- Положим ключ на SSD, а данные оставим на HDD
- Первой частью запроса с SSD полный ключ
- Второй – идти в HDD с найденными ключами на чтение данных



Попытка включения 3

Попытка включения 3

- Все заработало

Попытка включения 3

- Все заработало
- Но есть нюанс

Попытка включения 3

- Все заработало
- Но есть нюанс
- У нас большой поток на запись: 20gbit/sec

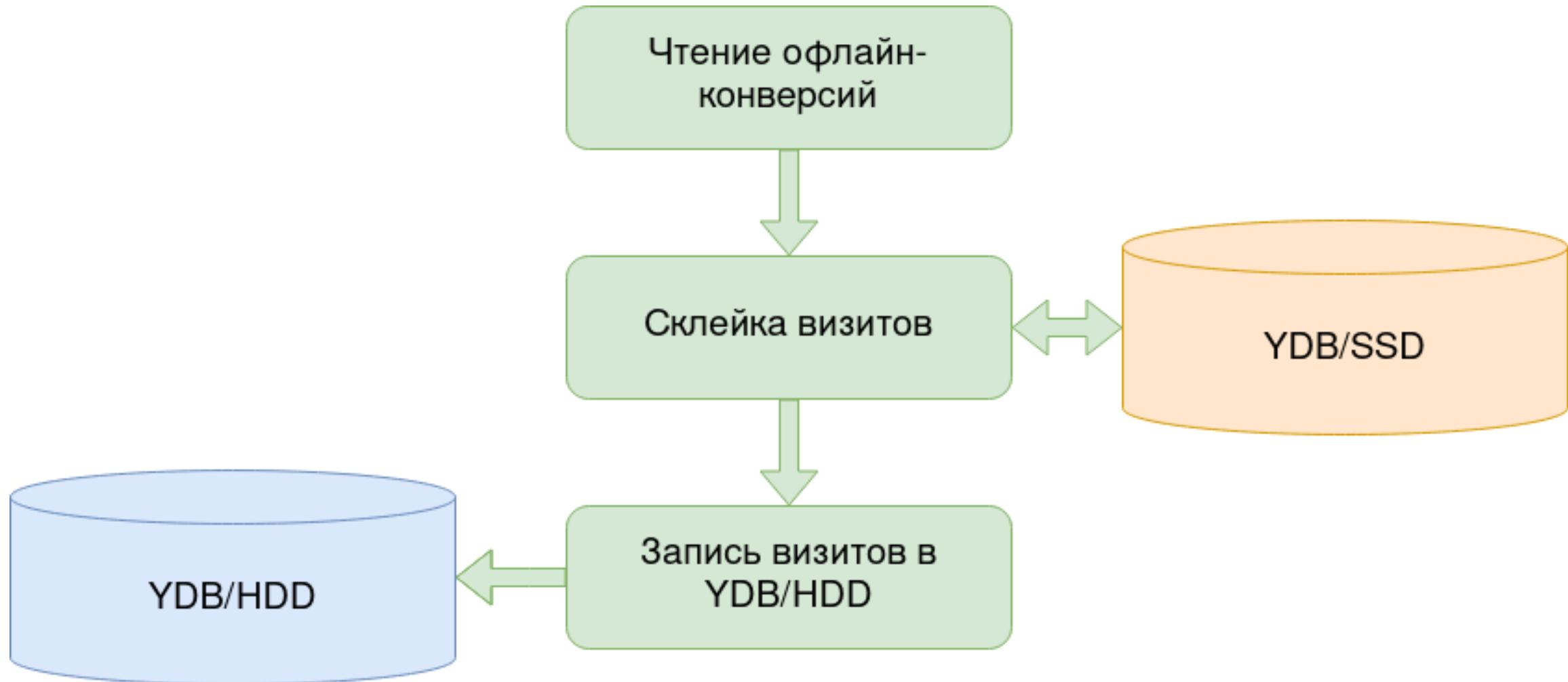
Попытка включения 3

- Все заработало
- Но есть нюанс
- У нас большой поток на запись: 20gbit/sec
- Значит критично писать без отставаний

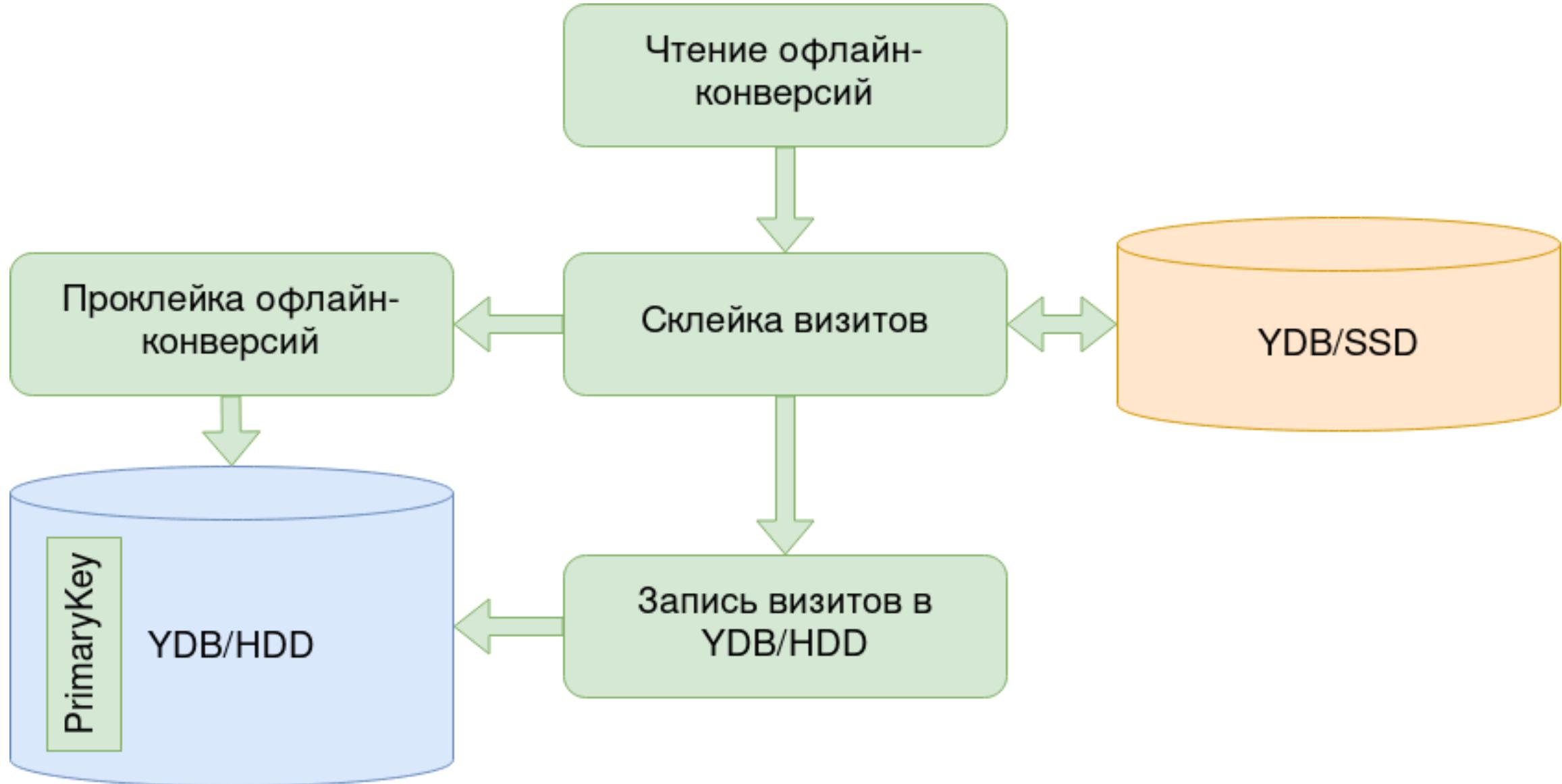
Попытка включения 3

- Все заработало
- Но есть нюанс
- У нас большой поток на запись: 20gbit/sec
- Значит критично писать без отставаний
- Чтение может задержаться на 5-10 минут

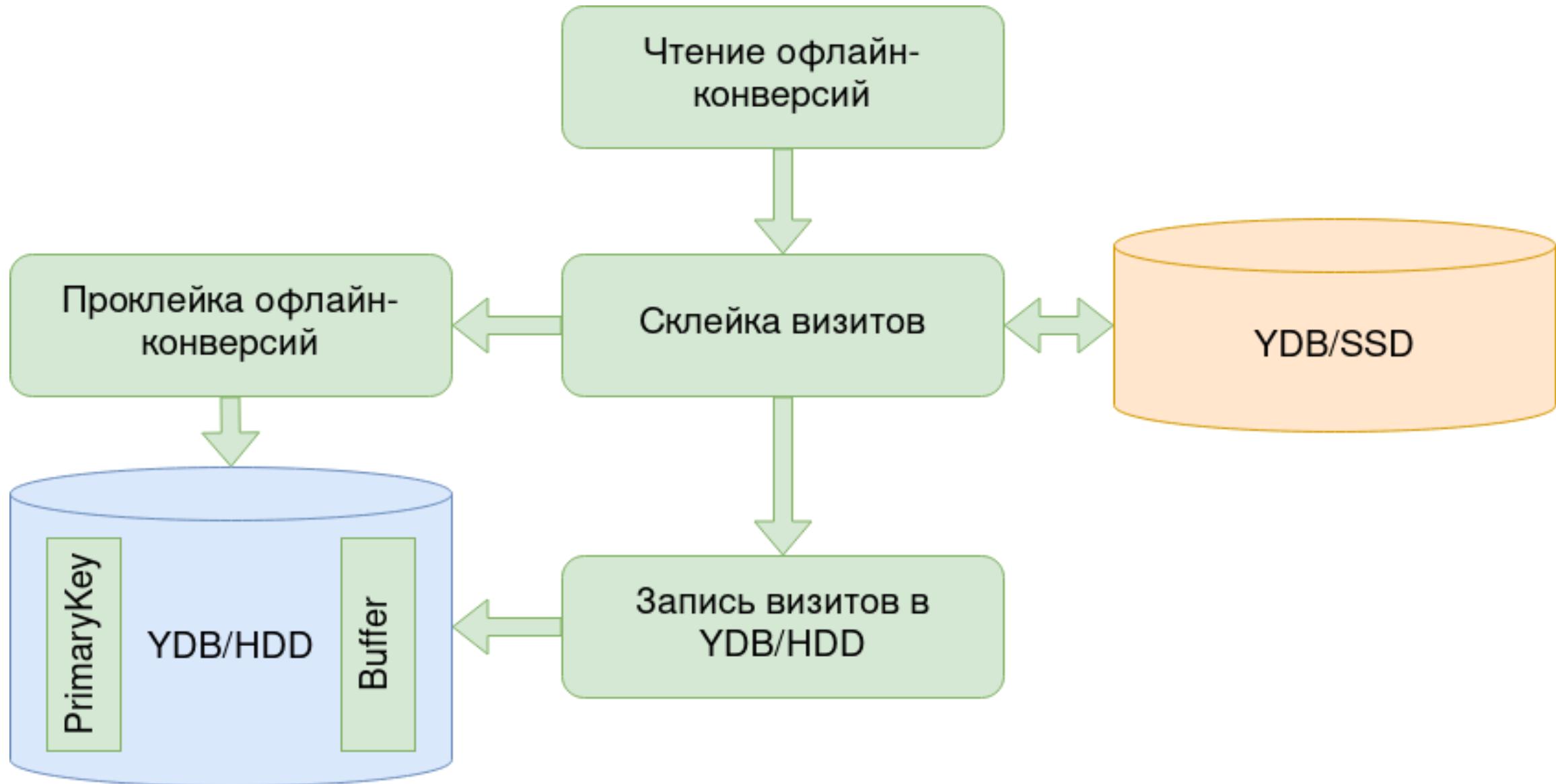
Первая версия схемы



Вторая версия схемы



Вторая версия схемы



Через 3 месяца после запуска

Через 3 месяца после запуска



Через 3 месяца после запуска

- Не забывайте выставлять правильно количество партиций

Через 3 месяца после запуска

- Не забывайте выставлять правильно количество партиций
- Автосплит работает, только если выставить min/max

Через 3 месяца после запуска

- Не забывайте выставлять правильно количество партиций
- Автосплит работает, только если выставить min/max
- Если не выставить uniform, начнет заливаться с 0



Через 3 месяца после запуска

- Не забывайте выставлять правильно количество партиций
- Автосплит работает, только если выставить min/max
- Если не выставить uniform, начнет заливаться с 0
- Если так сделать на HDD с дневными таблицами, то базе будет плохо 2-3 часа каждый день из-за compaction



Через 3 месяца после запуска

- Не забывайте выставлять правильно количество партиций
- Автосплит работает, только если выставить min/max
- Если не выставить uniform, начнет заливаться с 0
- Если так сделать на HDD с дневными таблицами, то базе будет плохо 2-3 часа каждый день из-за compaction
- Личная рекомендация: разброс между min/max около 20%



Что в итоге

- Залито 500 терабайт

Что в итоге

- Залито 500 терабайт
- В горизонте года – петабайт

Что в итоге

- Залито 500 терабайт
- В горизонте года – петабайт
- Общий поток записи сейчас 20 gbit/sec

Что в итоге

- Залито 500 терабайт
- В горизонте года – петабайт
- Общий поток записи сейчас 20 gbit/sec
- 300 дисков в инсталляции

Что в итоге

- Залито 500 терабайт
- В горизонте года – петабайт
- Общий поток записи сейчас 20 gbit/sec
- 300 дисков в инсталляции
- Уносим на HDD данные там, где редкие чтения и нет обязательных требований к очень низким задержкам



Что в итоге

- Залито 500 терабайт
- В горизонте года – петабайт
- Общий поток записи сейчас 20 gbit/sec
- 300 дисков в инсталляции
- Уносим на HDD данные там, где редкие чтения и нет обязательных требований к очень низким задержкам
- Надеемся, что YDB улучшится в части CPU+RAM

Вопросы?



Антон Барабанов
beckpost@yandex.ru



HighLoad ++