TODO Оптимизация join в ClickHouse

данная статья будет пересказом статьи https://clickhouse.com/docs/en/sql-reference/statements/select/join/ в части оптимизаций + некоторое исследование

Порядок таблиц при джойне

```
SELECT count()
FROM analytics.new_app_funnel_table AS wf
INNER JOIN analytics.int_spree_orders AS o ON wf.context_order_id = o.number
```

Правая таблица должна быть всегда меньше левой!

При джойне в оперативную память берется ПРАВАЯ таблица, если ее размеры выходят за определенный лимит, то алгоритм частично меняется на merge-sort join,

а merge-sort работает дольше чем hash join а если правая таблица очень большая то и merge-sort будет работать медленно

Фильтры

Все фильтры применяются после джойна, никаких оптимизаций здесь нет

```
SELECT count()
FROM analytics.new_app_funnel_table AS wf
INNER JOIN analytics.int_spree_orders AS o ON wf.context_order_id = o.number
WHERE toDate(created_at) = '2022-10-10'
```

здесь сначала произойдет джойн а потом фильтрация

Как правильно обращаться к таблицам при джойне

Обращаться к таблицам нужно через подзапросы

```
SELECT count()
FROM analytics.new_app_funnel_table AS wf
INNER JOIN analytics.int_spree_orders AS o ON wf.context_order_id = o.number

Elapsed: 2259.573 sec.

SELECT count()
FROM analytics.new_app_funnel_table AS wf
INNER JOIN
(
    SELECT *
    FROM analytics.int_spree_orders
    WHERE number IN (
        SELECT context_order_id
        FROM analytics.new_app_funnel_table
    )
) AS o ON wf.context_order_id = o.number

Elapsed: 725.149 sec. (wow!)
```

Здесь нужно обратить внимание на несколько моментов

1) операция IN в подзапросе по ключу джойна

2) наличие фильтра в подзапросе, туда нужно стараться добавлять как можно больше фильтров

3) в select в подзапросе нужно ставить только НУЖНЫЕ колонки здесь я взял * для теста

В идеале ваш запрос должен выглядеть как-то так

```
SELECT *
FROM
(
    SELECT
        context_order_id,
        appsflyer_id
    FROM analytics.new_app_funnel_table
    WHERE observation_date = '2022-10-10'
) AS wf
INNER JOIN
(
    SELECT number
    FROM analytics.int_spree_orders
    WHERE (number IN (
        SELECT context_order_id
        FROM analytics.new_app_funnel_table
    )) AND (toDate(created_at) = '2022-10-10')
) AS o ON wf.context_order_id = o.number

Elapsed: 69.167 sec.
```

Когда стоит применять вложенные запросы?

Всегда, когда возможно отфильтровать данные из таблицы перед джоином

С использованием IN в подзапросе нужно быть аккуратнее так как если две таблицы почти полностью пересекаются по ключу джоина то отфильтруется мало данных и запрос может стать наоборот медленнее

В приведенном выше примере я рассуждал так

```
SELECT countDistinct(context_order_id)
FROM analytics.new_app_funnel_table

uniqExact(context_order_id)
                    56198397


SELECT countDistinct(number)
FROM analytics.int_spree_orders

uniqExact(number)
         223075186


SELECT countDistinct(number)
FROM analytics.int_spree_orders
WHERE number IN (
    SELECT DISTINCT context_order_id
    FROM analytics.new_app_funnel_table
)

uniqExact(number)
          53562059
```

Получается в. этом случае удалось отфильтровать примерно 75% данных