

ПРОЕКТНОЕ ЗАДАНИЕ № 2.1

Удаление дублей из витрины dm.client

Условие:

После формирования отчетов заказчик обратил внимание на наличие дублирующихся строк в витрине dm.client.

Цель:

Устранить дубликаты так, чтобы в таблице осталась только одна уникальная запись для каждой пары:

- client_rk - уникальный идентификатор клиента;
- effective_from_date - дата начала действия записи.

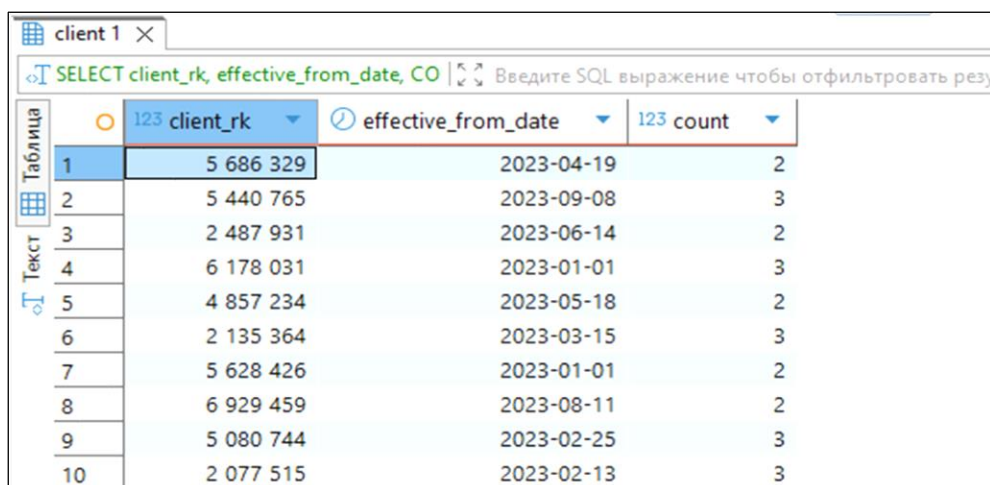
1. Поиск дублей:

```
SELECT client_rk, effective_from_date, COUNT(*)  
FROM dm.client  
GROUP BY client_rk, effective_from_date  
HAVING COUNT(*) > 1;
```

Описание:

Запрос находит группы записей, у которых совпадают client_rk и effective_from_date, и количество таких строк превышает одну - это и есть дубликаты.

Результат запроса поиска дублей:



	123 client_rk	effective_from_date	123 count
1	5 686 329	2023-04-19	2
2	5 440 765	2023-09-08	3
3	2 487 931	2023-06-14	2
4	6 178 031	2023-01-01	3
5	4 857 234	2023-05-18	2
6	2 135 364	2023-03-15	3
7	5 628 426	2023-01-01	2
8	6 929 459	2023-08-11	2
9	5 080 744	2023-02-25	3
10	2 077 515	2023-02-13	3

2. Проверка строк, подлежащих удалению:

```
SELECT *
FROM dm.client c
WHERE EXISTS (
  -- Проверяем наличие дубликатов (строка с таким же client_rk и
  effective_from_date) для текущей строки

  SELECT 1
  FROM dm.client sub
  WHERE sub.client_rk = c.client_rk
        AND sub.effective_from_date = c.effective_from_date
  -- Исключаем строку с минимальным ctid, считающуюся "оригиналом"
  -- Таким образом, выбираются только дубликаты, отличные от оригинала

        AND ctid <> (
  -- Определяем минимальный ctid для группы с одинаковыми client_rk и
  effective_from_date
  -- Минимальный ctid будет считаться оригиналом записи
  SELECT MIN(ctid)
  FROM dm.client
  WHERE client_rk = sub.client_rk
        AND effective_from_date = sub.effective_from_date)
);
```

Описание:

Запрос выводит все строки-дубликаты, кроме одной (оригинальной), которая имеет минимальный ctid (уникальный системный идентификатор строки в PostgreSQL) в каждой группе. Это позволяет убедиться, что удаляться будут только лишние копии записей.

Результат запроса плана удаления:

client 1						
SELECT * FROM dm.client c WHERE EXISTS: Введите SQL выражение чтобы отфильтровать результаты						
	123 client_rk	effective_from_date	effective_to_date	123 account_rk	123 address_rk	123 department_rk
1	3 284 359	2023-01-01	2023-01-19	2 592 554	308 208	557
2	3 284 359	2023-01-01	2023-01-19	2 592 554	308 208	557
3	3 284 359	2023-01-19	2023-02-25	5 518 658	916 056	557
4	3 284 359	2023-01-19	2023-02-25	5 518 658	916 056	557
5	3 284 359	2023-01-19	2023-02-25	5 518 658	916 056	557
6	3 284 359	2023-02-25	2999-12-31	3 934 952	515 668	557
7	3 284 359	2023-02-25	2999-12-31	3 934 952	515 668	557
8	7 089 917	2023-01-01	2023-01-28	2 107 352	735 745	748
9	7 089 917	2023-01-01	2023-01-28	2 107 352	735 745	748

3. Удаление дубликатов:

```
DELETE FROM dm.client
WHERE ctid NOT IN (
-- Оставляем только одну (первую) строку в каждой группе дубликатов
-- Группируем по составному ключу: client_rk и effective_from_date
-- MIN(ctid) выбирает "оригинальную" строку (первая добавленная в БД)

    SELECT MIN(ctid)
    FROM dm.client
    GROUP BY client_rk, effective_from_date
);
```

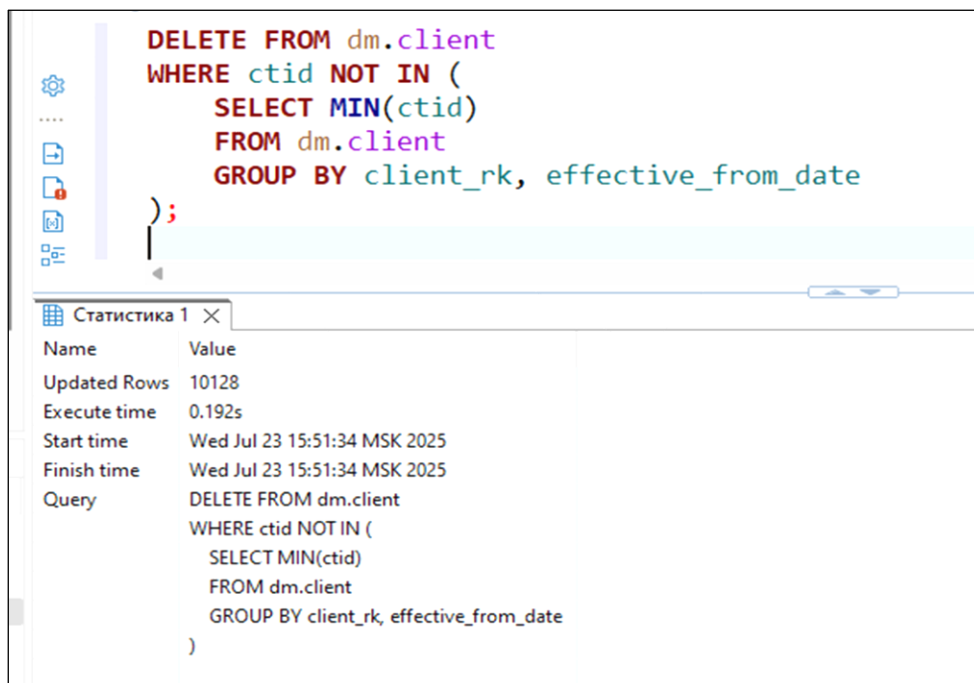
Описание:

Удаляются все строки, не являющиеся "оригинальными" - то есть, у которых ctid **не** является минимальным в группе с одинаковыми client_rk и effective_from_date.

Пояснение по ctid:

ctid - это системное поле PostgreSQL, указывающее на физическое местоположение строки в таблице. Оно уникально в пределах таблицы и может использоваться для удаления дубликатов, когда отсутствует уникальный первичный ключ.

Результат запроса удаления:



The screenshot shows a PostgreSQL query editor with a DELETE statement and its execution statistics. The query is as follows:

```
DELETE FROM dm.client
WHERE ctid NOT IN (
    SELECT MIN(ctid)
    FROM dm.client
    GROUP BY client_rk, effective_from_date
);
```

The statistics window, titled "Статистика 1", displays the following information:

Name	Value
Updated Rows	10128
Execute time	0.192s
Start time	Wed Jul 23 15:51:34 MSK 2025
Finish time	Wed Jul 23 15:51:34 MSK 2025
Query	DELETE FROM dm.client WHERE ctid NOT IN (SELECT MIN(ctid) FROM dm.client GROUP BY client_rk, effective_from_date)

4. Проверка результата:

Повторно выполняем запрос на поиск дубликатов, чтобы убедиться, что в таблице осталась только одна запись на каждую уникальную пару client_rk и effective_from_date.

Результат запроса поиска дублей после удаления:

...

↓

↓

↓

↓

```
SELECT client_rk, effective_from_date, COUNT(*)
FROM dm.client
GROUP BY client_rk, effective_from_date
HAVING COUNT(*) > 1;
```

client 1 ×

SELECT client_rk, effective_from_date, CO

Введите SQL выражение чтобы отфильтровать резул

123 client_rk

effective_from_date

123 count

Таблица

Текст