

 Разработка NTA  16.06.2020

## Способы удаления дубликатов в SQL Server

При проектировании объектов, в частности таблиц в БД SQL Server, необходимо придерживаться определенных правил. Однако, даже если следовать данным правилам существует вероятность появления дубликатов в строках таблиц. Данная статья посвящена различным способам очистки данных от дубликатов.

 1 

7219 просмотров

При проектировании объектов, в частности таблиц в БД SQL Server необходимо придерживаться определенных правил: рекомендуется использовать правила нормализации БД; таблица должна иметь первичные ключи, кластерные и некластерные индексы; ограничения для обеспечения целостности данных и производительности. Но даже если следовать этим правилам, мы можем столкнуться с проблемой появления дубликатов в строках таблицы. Кроме этого, возможна ситуация получения дубликатов при импорте данных, когда мы загружаем данные as is в промежуточные таблицы, и далее требуется удалить дублирующие записи перед загрузкой в промышленные таблицы.

Рассмотрим различные способы для очистки данных от дублей. Создадим простую таблицу сотрудников и наполним её несколькими записями.

```
CREATE TABLE Employee
(
    [id]                int identity(1,1),
    [Фамилия]           nvarchar(100),
    [Имя]               nvarchar(100),
    [Отчество]          nvarchar(100),
    [Дата рождения]    date,
)
GO

Insert into Employee ([Фамилия],[Имя],[Отчество],[Дата рождения])
values
(N'Алексеев',N'Алексей',N'Алексеевич','1990-03-01'),
(N'Алексеев',N'Алексей',N'Алексеевич','1990-03-01'),
(N'Алексеев',N'Алексей',N'Алексеевич','1990-03-01')
```



```
(N'Иванов',N'Иван',N'Иванович','1985-01-01'),  
  
(N'Иванов',N'Иван',N'Иванович','1985-01-01'),  
(N'Петров',N'Петр',N'Петрович','1988-02-01'),
```

Как мы видим, в таблице присутствуют дублирующие строки, которые необходимо удалить.

- Удаление дубликатов с использованием агрегатных функций

С помощью условия GROUP BY мы группируем данные по определенным столбцам и используем функцию COUNT для подсчета вхождений строк в таблицу.

Например, с помощью следующего запроса, определим записи, которые присутствуют в таблице более 1 раза.

```
Select [Фамилия], [Имя], [Отчество], [Дата рождения], count(*) as CNT  
FROM NTA.dbo.Employee  
GROUP BY [Фамилия], [Имя], [Отчество], [Дата рождения]  
having count(*) > 1
```

Т.е. сотрудники Алексеев А.А. и Иванов И.И. присутствуют в таблице 3 и 2 раза соответственно.

	Фамилия	Имя	Отчество	Дата рождения	CNT
1	Алексеев	Алексей	Алексеевич	1990-03-01	3
2	Иванов	Иван	Иванович	1985-01-01	2

Удалим дублирующие записи, оставив только строки с MIN id сотрудника.

```
Delete FROM NTA.dbo.Employee  
Where id not in  
(  
    select min(id) as MinRowID  
    FROM NTA.dbo.Employee  
    group by [Фамилия],[Имя],[Отчество],[Дата рождения]  
)
```

Выведем оставшиеся записи таблицы, и убедимся, что дубликаты отсутств,

Отметим, что данный способ удаления дубликатов возможен в случае таблиц, для которых определен первичный ключ.

	id	Фамилия	Имя	Отчество	Дата рождения
1	1	Алексеев	Алексей	Алексеевич	1990-03-01
2	4	Иванов	Иван	Иванович	1985-01-01
3	6	Петров	Петр	Петрович	1988-02-01

- Удаление дубликатов используя обобщенные табличные выражения (CTE)

Мы можем использовать связку обобщенных табличных выражений и функции ROW\_number() для удаления дубликатов, например следующим образом:

```
WITH CTE ([Фамилия],
          [Имя],
          [Отчество],
          [Дата рождения],
          [Нумерация]
)
AS (SELECT      [Фамилия],
               [Имя],
               [Отчество],
               [Дата рождения],
               ROW_NUMBER () OVER (PARTITION BY [Фамилия],
                                               [Имя],
                                               [Отчество],
                                               [Дата рождения]
                                   ORDER BY id) AS [Нумерация]

    FROM NTA.dbo.Employee)
DELETE FROM CTE
WHERE [Нумерация] > 1
```

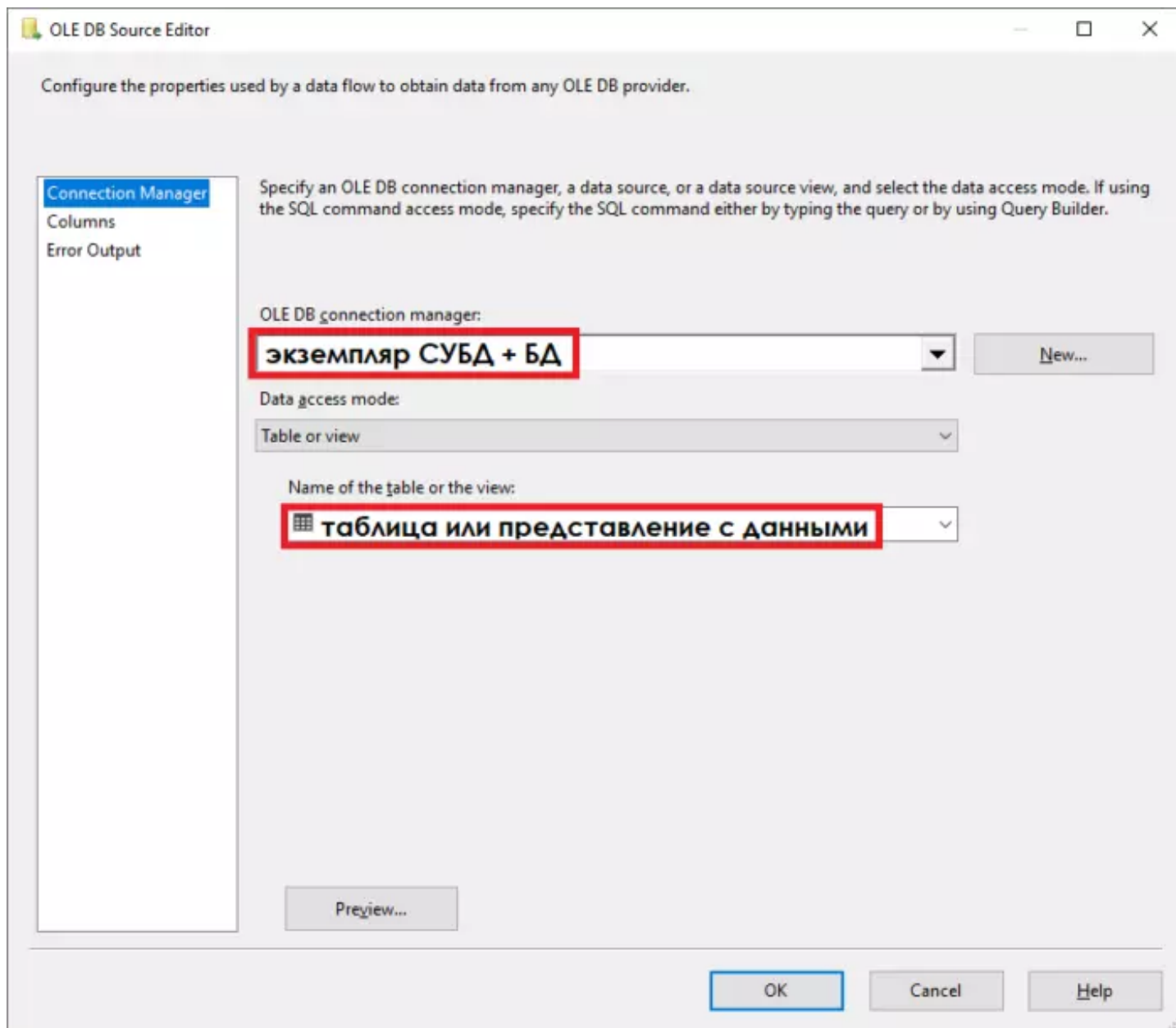
В данном запросе мы используем функцию ROW\_number() с конструкцией partition BY в предложении OVER для нумерации записей, и удаляем записи с пронумерованными значениями > 1, соответствующие дубликатам.

- Удаление дубликатов с использованием инструментария SSIS пакетов.

Создадим в SQL Server Data Tools новый пакет integration Services.

Добавим в пакет элемент «OLE DB Source», откроем редактор OLE DB Source, в графе Connection Manager укажем реквизиты экземпляра СУБД и БД, и наименование исходной таблицы с данными, содержащей дубликаты.





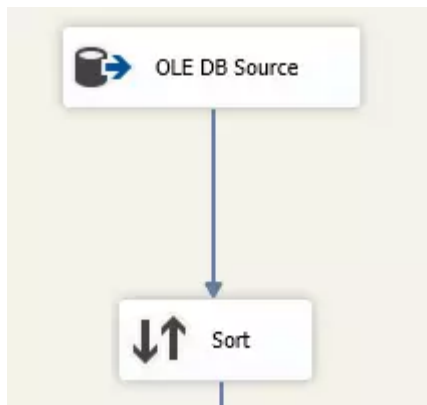
С помощью кнопки Preview убедимся, что в исходной таблице присутствуют дубликаты.

Preview Query Results

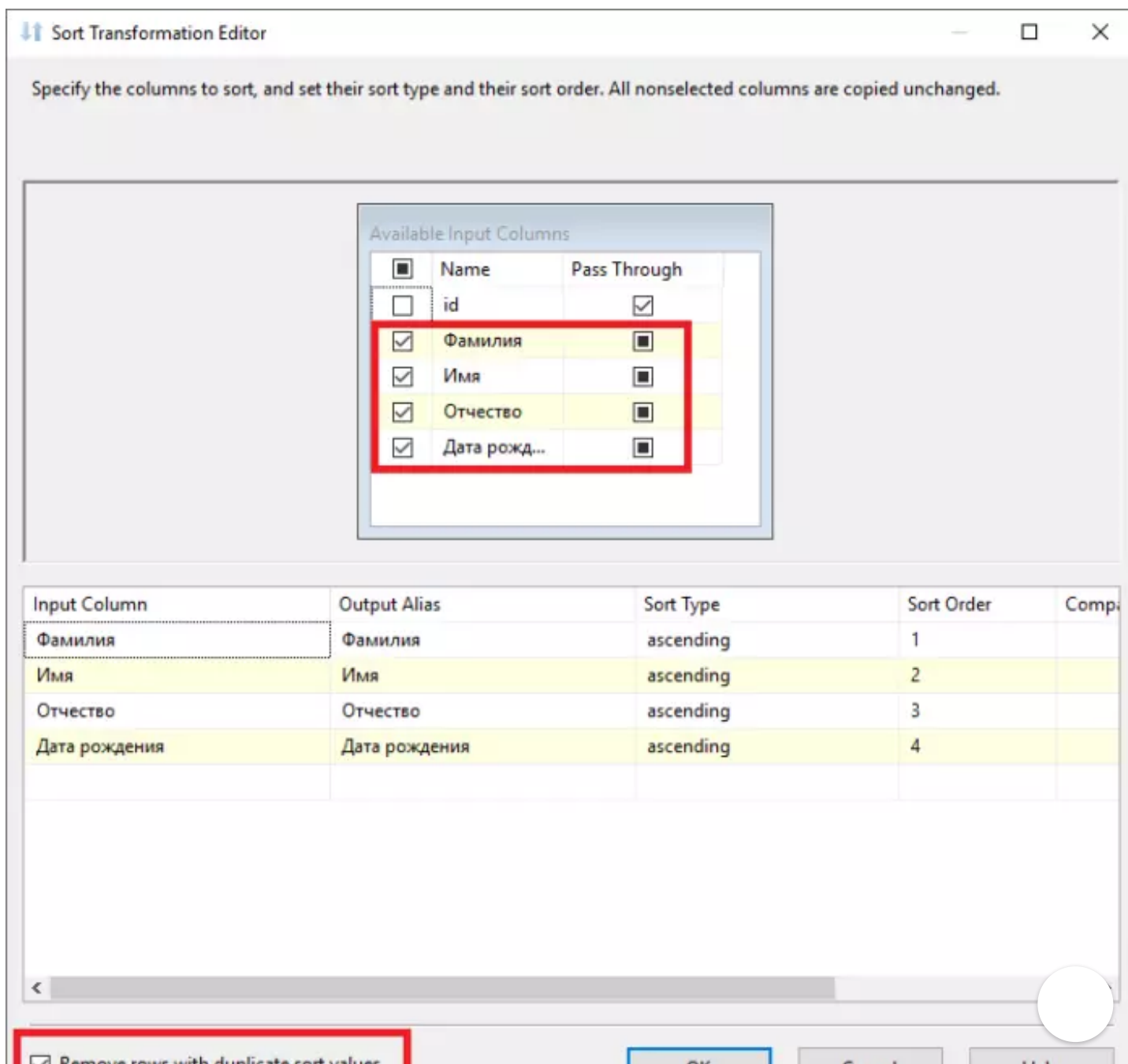
Query result (up to the first 200 rows):

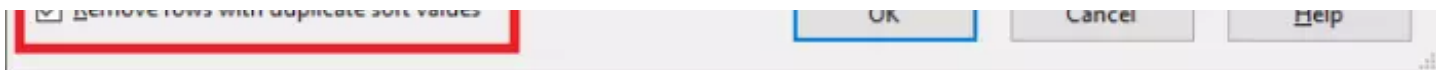
id	Фамилия	Имя	Отчество	Дата рож...
1	Алексеев	Алексей	Алексеевич	1990-03-01
2	Алексеев	Алексей	Алексеевич	1990-03-01
3	Алексеев	Алексей	Алексеевич	1990-03-01
4	Иванов	Иван	Иванович	1985-01-01
5	Иванов	Иван	Иванович	1985-01-01
6	Петров	Петр	Петрович	1988-02-01

добавим оператор «Sort», и выделим поля, в которых присутствуют дублирующие данные.

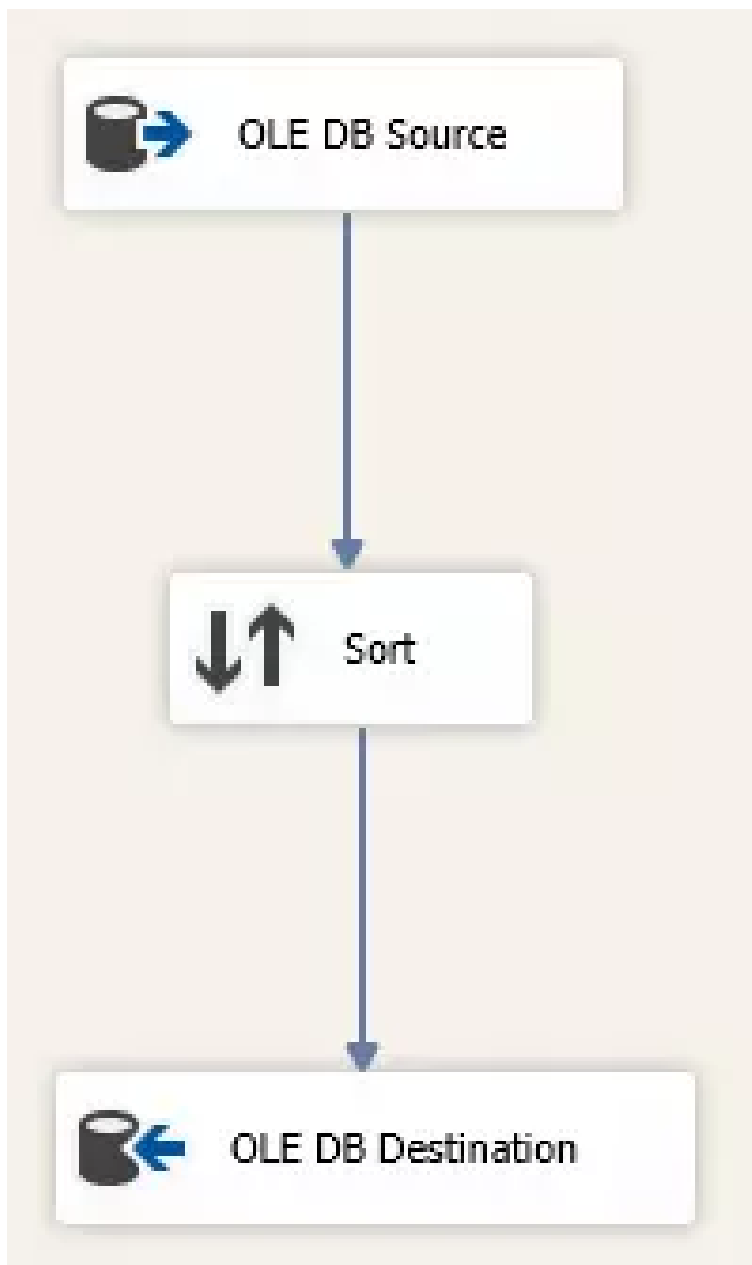


Установим галку «Remove rows with duplicate sort values» для удаления дубликатов.



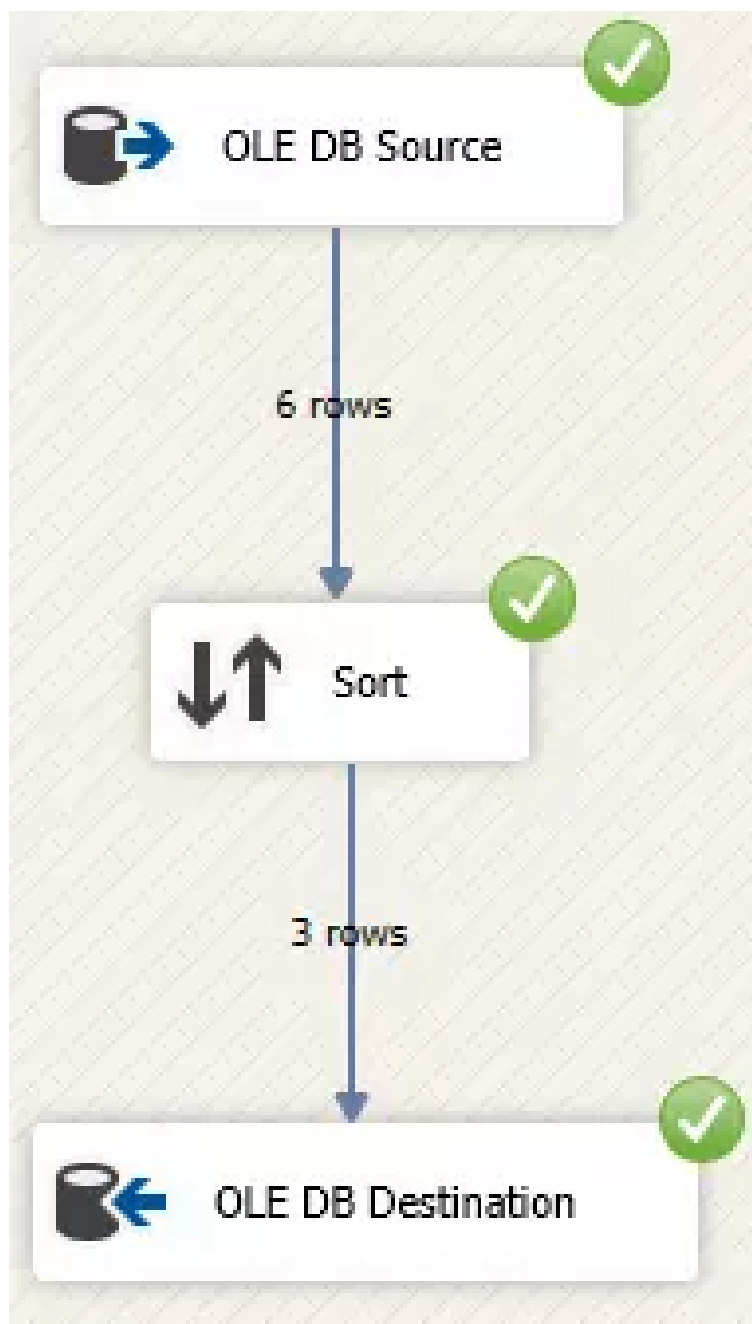


Добавим элемент «OLE DB Destination», в котором укажем целевую таблицу для записей результата очистки данных.



Запустив на исполнение реализованный SSIS пакет, мы видим, что в целевой источник загрузилось 3 строки, проверим, что отсутствуют дубликаты.





	id	Фамилия	Имя	Отчество	Дата рождения
1	3	Алексеев	Алексей	Алексеевич	1990-03-01
2	5	Иванов	Иван	Иванович	1985-01-01
3	6	Петров	Петр	Петрович	1988-02-01

Необходимо отметить, что при использовании данного способа потребуется дополнительное место для хранения новой целевой таблицы, однако данный вариант позволяет избежать ошибок и вернуться к исходному варианту, в случае если результат в целевой таблице не будет являться удовлетворительным.



В данной статье мы рассмотрели различные способы удаления дубликатов записей в таблицах БД SQL Server, которые могут быть использованы в работе в зависимости от задачи и объема данных.

При больших объемах дубликатов в данных целесообразно рассмотреть возможность сохранения уникальных значений в промежуточную таблицу, очистку рабочей таблицы, и возврат оставленных уникальных записей.

1

1

NTA

Лайфхаки IT, проверенные решения для стандартных задач

Подписаться

Создать объявление

Отключить

Готовые таргетинги для Instagram

50 тысяч вариантов детального таргетинга, которые не нужно подбирать ночами

Получить

Вакансии

Разместить

Backend Developer (PHP)

Uma.Tech Удалённо

Стажировка: IT-редактор

Glyph Media Удалённо от 15 000 до 30 000 ₽

Scrum-мастер, Agile Coach

Точка Екатерин... от 100 000 до 250 000 ₽

Менеджер, руководитель отдела интернет-продаж (b2c)

Bonanza Москва от 80 000 до 150 000 ₽

Head of Performance Marketing

Mirafox Удалённо от 150 000 ₽



Показать ещё ▾

1 комментарий

Популярные По порядку



Prolis Labkk

16 июн 2020

Delete FROM NTA.dbo.Employee Where id not in  
- плохое начало хорошего дня.

🗨 Ответить 📌 ...

▾ 0 ▴

Написать комментарий...



Отправить

Мероприятия


Разместить

Показать ещё ▾

Блоги компаний




## Как я на коленке делал свой бесплатный курс программирования, о котором мечтал 3 года

В этой статье я расскажу о MVP курса по обучению программированию, который возник благодаря удаленке. 




## Несколько забавных, и не только, фишек Python

Каким бы весёлым (или скучным?) не казалось вам программирование, всегда можно внести в этот процесс ещё больше красок.... 



## Елена Зленко: «Москва постепенно превращается в экомегополис»

Сегодня у нас есть возможность получать не только исчерпывающую информацию о состоянии окружающей среды нашего города,... 

Показать еще 

### Лучшие комментарии

☐ Павел Бондаренко

131

Если бы да кабы - во рту выросли грибы - то был бы не рот - а полный огород. Если бы он в 2011 купил на 43000\$ биткоинов, то сейчас бы продал их за 100500 миллионов. Точно так же мог бы тогда инвестировать в S&P - и...

Инвестиции в однушки в РФ: итоги 10 лет

☐ Gidevan

338

Ну по фотографиям не понятно был ли пожар. Тут без эксперта не обойтись.

«Тинькофф.Страхование»: экспертизу назначили после того, как мы по согласованию вынесли сгоревшую мебель из квартиры

### Еженедельная рассылка

Одно письмо с лучшим за неделю

 Почта



