Search
Write
Sign in
Get unlimited access to the best of Medium for less than $5/week. Become a member

Data Engineering

Deleting Duplicate Data With No Primary Key — A Data Engineer's Worst Nightmare?

Intermediate SQL for ETL Dev to Data Engineer Transition

Saikat Dutta · Follow
Published in Data Engineer Things · 5 min read · Mar 13, 2024

150 2

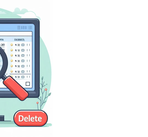


Image generated by Bing using Dalle.

Imagine, and be everyday in the nearby buildings is fast asleep. But you are not able to sleep. You keep tossing and turning in your bed, praying the job doesn't break today.

It's your turn for the Call, one of the worst responsibilities of an ETL Developer or a Data Engineer.

The On-Call Nightmare, A failed job:

Suddenly the phone rings, it's a call from the support operations team of an urgent issue reported by the business. Their dashboard is not returning data within SLA.

When we start praying for a quiet day, the world seems to have different plans.

You jump out of your bed in pajamas, rush to your office setup and boot your laptop, still sweating and struggling to catch your breath.

You are not sure, what might have caused the report to slow down. You think to yourself, is there a huge data set that has come in?

And then, your worst nightmare comes to life. The destination table does not have any primary key. A list of duplicate rows have made it to the destination table.

The duplicate records are causing the jobs to fail, and pull a lot of unintended data.

The junior support associate is waiting for your instructions. But the question is:

How do you identify how many duplicate records are there?

How do you delete only the duplicate ones we intend to delete?

Let's find out ….

Setup Sample Data

Let's try to demonstrate the above scenario with some very basic data. We will create a table employee, and insert some duplicate and correct records in the table.

```
create table employee

(
empid int,
firstname varchar(100),
lastname varchar(100)
);

insert into employee values(1,'Jack','1969-09-0000');
insert into employee values (2, 'Jim', '1969-09-0000');
insert into employee values (3, 'Jack', '1969-09-0000');
insert into employee values (4, 'Mike', '1969-09-0000');
insert into employee values (5, 'Jim', '1969-09-0000');
insert into employee values (6, 'Jim', '1969-09-0000');
insert into employee values (7, 'Jack', '1969-09-0000');
insert into employee values (8, 'Mike', '1969-09-0000');
insert into employee values (9, 'Jim', '1969-09-0000');
insert into employee values (10, 'Jack', '1969-09-0000');
insert into employee values (11, 'Mike', '1969-09-0000');
insert into employee values (12, 'Jim', '1969-09-0000');

select * from employee
```

Duplicate data in the **employee** table

As you can see, Jack, Mike, Cathy and Lisa, all have duplicate records in all the columns in the table now. We can see this on our own with such a small data set.

However, there must be a better way to identify the duplicate records.

How to find duplicate data:

To identify duplicate data, group by or ROW_NUMBER() is the preferred response. To identify a specific set of records as duplicate, we need to identify what is the candidate to be the primary key of the employee table, which should be unique.

In this case, let's consider Empid, Emphone and Empfirst as the key columns. The below query takes a count of records in the table, corresponding to each group of Empid, Emphone and Empfirst.

```
SELECT Empid, Emphone, Empfirst, count(1) as no_of_dupes
FROM employee
group by Empid, Emphone, Empfirst
```

Number of duplicate entries

As we can see except for Joe and Fred, all the other employees have their data duplicated.

We could also use the ROW_NUMBER() function in MS SQL. Query and results below:

```
SELECT ROW_NUMBER() OVER (PARTITION BY Empid, Emphone, Empfirst ORDER BY Empid) rn
FROM Employee
```

ROW_NUMBER() to identify duplicates.

Deleting duplicate data:

Ok, we have identified the duplicate records, however, how can we get rid of them? Thankfully there are ways we can do this:

Using ROW_NUMBER():

```
with rnkcounter as (SELECT ROW_NUMBER() OVER (PARTITION BY Empid, Emphone, Empfi
select *
delete from rnkcounter
where rnk > 1;
```

The above query deletes 6 rows of duplicate data and leaves only one record, per batch.

In this SQL statement:

- The CTE uses the ROW_NUMBER() function to find the duplicate rows partitioned by values in the Empid, Emphone, and Empfirst columns.
- Then, the DELETE statement deletes all the duplicate rows but keeps only one record per each duplicate batch.

What are the other ways to do this?

We can use many different approaches to do the same. Here are some worth exploring:

Use TOP statement to delete all records except only one per id passed in the variable qtid. This is not the best method as the queries will need to be repeated for each Empid.

```
DECLARE qtid INT = 3

DELETE TOP (SELECT COUNTD(1) -1) FROM dbo.Employee WHERE Empid = qtid;
FROM dbo.duplicateTable
WHERE id = qtid
```

Use a backup table along with distinct columns to load only the distinct columns into a new table. Distinct function will only return one record for every duplicate record. Drop the table with duplicates. Rename the new table as the main **employee** table.

```
-- Use select distinct to get only distinct records into a backup table
select distinct *
into dbo.Employee_bkp
from dbo.Employee

select * from dbo.Employee_bkp

-- Drop the original table with dupes.
drop table dbo.Employee

-- Rename the new table as the employee
EXEC sp_rename 'dbo.Employee_bkp', 'employee'

select * from dbo.Employee
```

Warning:

Though we had all the column duplicated in this case, in an actual production environment we might see only a subset of columns duplicated.

1. If the duplicate columns qualify as valid primary key columns, then we can use the approach to only use those duplicate columns in the ROW_NUMBER() function.
2. Always take a backup of the main table before deleting any data
3. Don't forget to test your queries in the non-production environment with dummy data.
4. A delete statement should always accompany a where clause to limit the number of records.
5. Be extremely careful before running any delete statement in the production environment.

Read about Self Join here.

Whenever you're ready, there are 3 ways I can help you in your Career Growth:

1. Let me help you mentor with your career journey here.
2. Grow Your LinkedIn brand and get diverse here.
3. Take charge of your career growth here.

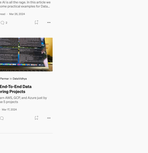
Book some time with me: https://topmate.io/saikatdutta

Sql · Data Engineering · Data Engineering SQL · Interview · Sqlinterviewtips

150 2

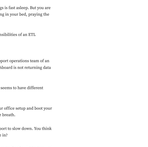
Written by Saikat Dutta
234 Followers · Writer for Data Engineer Things
Follow
Azure Data Engineer|Multi Cloud Data Professional|Data Architect |Career Mentor | Writer|Mail:-Saikatdutta.dataengineer@GMAIL.com

More from Saikat Dutta and Data Engineer Things

Saikat Dutta in Towards Data Science
When Do You Self Join? A Handy Trick
Intermediate SQL for ETL dev to Data Engineer Transition
7 min read · Mar 21, 2024
Hussein Jundi in Data Engineer Things
Data Engineering Architectures & Strategies for Handling Sensitive...
Data Engineering Architectures & Strategies for Handling Sensitive Data Efficiently and securely...
7 min read · Mar 2, 2024

Saikat Dutta in Data Engineer Things
I spent 7 hours reading another paper to understand more about...
A data Cloud Developer/Lecturer looking at a death-spiral engine on Snowflake.
Mar 4, 2024
Saikat Dutta in Data Engineer Things
How to Create First Data Engineering Project? An...
Build Data Engineering projects in a fully Microsoft Azure environment for guaranteed...
Aug 20, 2023
See all from Saikat Dutta
See all from Data Engineer Things

Recommended from Medium

Akash Rajan
Data Engineering concepts Part 8: Data Architecture
Hi Folks, This is the 8th part series of Data Engineering concepts and in this part, we...
7 min read · Mar 30, 2024
Tarun Sethi · 3 Data Engineers
How I Built This Data Platform in One Week
Hey, I'm Tarun and I recently led my company's project (and the first) data engineering hire to release the entire...
7 min read · Mar 20, 2024

Stories to Help You Grow as a Software Developer
Bookmarks · 340 stories
Staff Picks
48 stories · 298 saves
Community Networks in Stackademic
Python Coding Questions for Data Engineer Interview Part 4 (Easy,...
Mastering Python For Data Engineering Interviews
4 min read · Jun 21, 2024
Vigen in Towards Data Science
How I use Gen AI as a Data Engineer
Generative AI is all the hype, in this article we will explore how I use Generative AI for Data Engineering.
8 min read · Mar 25, 2024

Jagadish Jangala
Tricky Scenario Question for Senior Data Engineers
Hi Folks, Recently I met a senior data engineer responsible for the company's sales data...
4 min read · Mar 7, 2024
David Freeman · 1 Stackoffice
5 FREE End-To-End Data Engineering Projects
Hey Data Engineers, Here are 5 free projects you can find doing these 5 projects
8 min read · Mar 15, 2024
See more recommendations
Help · Status · About · Careers · Blog · Privacy · Terms · Text to speech · Teams