

most efficient way of deleting ...



rmenon 57 posts since

Apr 23, 2009

Folks

For removing duplicate records from a table in Netezza, what is the most efficient way?

I am thinking something like:

```
delete from dupl
```

```
where rowid in
```

```
(
```

```
select a.rid
```

```
from
```

```
(
```

```
select rowid rid, row_number() over (partition by x,y,modification_ts order by rowid) rn
```

```
from dupl
```

```
) a
```

```
where rn != 1
```

```
);
```

Is this what people typically use or is there some other mechanism that is known to be more efficient? Thanx!

most efficient way of deleting ...



[Shawn Fox](#) 520 posts since

Aug 15, 2006 1. **Re: most efficient way of deleting duplicates in netezza** Feb 5, 2010 10:56 AM

The easiest way:

```
create table nodups as select distinct * from table_with_dups;
```

This may not work on really huge tables, but I've done it on a 1.5TB table on a 10400. It won't work if you have timestamp columns or something which might be different between two rows.



[rmenon](#) 57 posts since

Apr 23, 2009 2. **Re: most efficient way of deleting duplicates in netezza** Feb 5, 2010 10:59 AM

in response to: [Shawn Fox](#)

Thank - yeah a quick test reveals it to be quite a bit faster than the delete I proposed.

"It won't work if you have timestamp columns or something which might be different between two rows"

Did not quite understand the above statement => why won't it work when there are timestamps?



[Shawn Fox](#) 520 posts since

Aug 15, 2006 3. **Re: most efficient way of deleting duplicates in netezza** Feb 5, 2010 11:01 AM

in response to: [rmenon](#)

I meant it won't work if the timestamps are different or if you have any differences at all in the records. Sometimes you might have duplicates based off your primary key, but different timestamps or some other column like that.

most efficient way of deleting ...



[Jeff Feinsmith](#) 11 posts since

Mar 28, 2007 4. **Re: most efficient way of deleting duplicates in netezza** Feb 5, 2010 1:15 PM

in response to: [Shawn Fox](#)

To avoid the problem that Shawn describes, and perhaps a slightly faster approach, try this:

```
create new_table as select pk1, pk2, pk3, max(othercol1), max(othercol2), max(othercol3)
from old_table group by pk1, pk2, pk3;
```

This basically does the same thing as the distinct approach, but you get to choose your key columns. Everything else, including those pesky datetimes, gets rolled up into the max value. It should also be a bit faster because you only need to perform distincting on a subset of the columns.

-Jeff

Message was edited by: Jeff Feinsmith



[rmenon](#) 57 posts since

Apr 23, 2009 5. **Re: most efficient way of deleting duplicates in netezza** Feb 5, 2010 11:12 AM

in response to: [Shawn Fox](#)

"I meant it won't work if the timestamps are different or if you have any differences at all in the records. Sometimes you might have duplicates based off your primary key, but different timestamps or some other column like that."

I think you mean if the pk columns are the same and the time stamps (or some other column) are different somehow and if you want to consider such rows as duplicates, correct? If so, it makes sense to me.

most efficient way of deleting ...



[rmenon](#) 57 posts since

Apr 23, 2009 6. **Re: most efficient way of deleting duplicates in netezza** Feb 5, 2010 11:13 AM

in response to: [Jeff Feinsmith](#)

Thanx Jeff,

Interesting. That approach makes sense if you want to choose your "unique key" columns.



[Mark Aukeman](#) 24 posts since

Oct 14, 2008 7. **Re: most efficient way of deleting duplicates in netezza** Feb 5, 2010 12:47 PM

in response to: [Jeff Feinsmith](#)

This approach eliminates duplicates, but you need to be careful that the MAXed columns truly make business sense, like an arrival timestamp. If you MAX a code column or foreign key that varies within your grouped primary key columns, then the results may be inaccurate. The dupes are gone, but the problem is "swept under the carpet", so to speak. Here is a suggestion to avoid inaccurate results: 1) Confirm the true primary key, 2) ensure that all non-additive columns exist in the Group By except the time stamp 3) If there are still duplicates relative to the primary key, then get business criteria for which of the duplicates to preserve and which can be eliminated..



[Superuser](#) 90 posts since

Sep 19, 2008 8. **Re: most efficient way of deleting duplicates in netezza** Feb 6, 2010 12:58 AM

in response to: [Mark Aukeman](#)

oh, lot of approaches...

we use group by to achieve this

first check whether dups are there in the table according to the columns which are not supposed to have dups (excluding ur inserted/updated timestamp columns or something which differentiate ur dups)

```
select key1,key2,key3 from old_table group by key1,key2,key3 having count(1)>1
```

if yes, then

```
create new_table as select * from old_table where 1=2
```

most efficient way of deleting ...

```
insert into new_table(key1,key2,key3) select key1,key2,key3 from old_table group by  
key1,key2,key3
```