Eliminating Duplicates In a Query When Distinct is not an Option

Posted At : October 27, 2006 6:28 PM | Posted By : Mark Kruger
Related Categories: MS SQL Server

Here's a problem perhaps you have had. You want to select unique email addresses, first names and last names out of a database for a newsletter or to sell them a new mortgage or whatever. Being the nice guy that you are you don't want to send them multiple messages, so you want to eliminate duplicates, right? the problem is that *SELECT DISTINCT...* doesn't always work in this instance. For example, John Doe put his information in as John Doe in one case and John H. Doe in another. Selecting distinct for name and email will give you a duplicate name with the *same* email. Now obviously you could solve this problem in your Coldfusion code - but wouldn't it be nice to fix up the query?

This is exactly the problem we ran into recently. We had people signing up for "seminars" for a client and we would end up with duplications from folks who had signed up more than once. The fix? Join the table to itself, group by and use HAVING - like so:

```
<cfquery name="getEmails" datasource="#mydsn#">
    SELECT su.userid, su.seminarid as id, su.fname, su.lname, su.email
    FROM seminarusers su
    INNER JOIN seminarusers su2 ON su.uuid = su2.uuid
    WHERE su.seminarid IN (5,6)
    AND su2.seminarid IN (5,6)
    GROUP BY su.userid, su.seminarid, su.fname, su.lname, su.email
    HAVING su.userid = Max(su2.userid)
</cfquery>
```

In this example we joined the table to itself, grouped by the userid and then used having to get only the latest record. Very nifty! Thanks to CF Webtools developer Ryan Stille for the neat trick. One thing to note, the where clause has to apply to BOTH of the table aliases or you will get an unexpected result - so we did WHERE su.seminarid IN (5,6) AND su2.seminarid IN (5,6). You could just as easily remove duplicates this way (see my post on **Updating a table using a Join**).