

Subject: Sawmill Newsletter: Using Cross-Reference Tables
From: Jo Dee Koller <jodee@flowerfire.com>
Date: Fri, 12 Jun 2009 16:31:33 -0700
To: Greg Ferrar <ferrar@flowerfire.com>



Sawmill Newsletter

June 15, 2009

Welcome to the Sawmill Newsletter!

You're receiving this newsletter because during the downloading or purchase of Sawmill, you checked the box to join our mailing list. If you wish to be removed from this list, please send an email, with the subject line of "UNSUBSCRIBE" to newsletter@sawmill.net.

News

Sawmill 8.0.8 shipped on May 20, 2009. This is a minor "bug fix" release, and it is free to existing Sawmill 8 users. It is recommended for anyone who is experiencing problems with Sawmill 8.0.7 or earlier. You can download it from <http://sawmill.net/download.html>.

Sawmill 7 users can upgrade to Sawmill 8 for half of the license price; or if you have Premium Support, the upgrade is free. Major features of Sawmill 8 include support for Oracle and Microsoft SQL Server databases, real-time reporting, a completely redesigned web interface, better multi-processor and multi-core support, and role-based authentication control.

This issue of the Sawmill Newsletter describes the use of cross-reference tables to increase the speed of custom reports.

Get The Most Out Of Sawmill With Professional Services

Looking to get more out of your statistics from Sawmill? Running short on time, but need the information now to make critical business decisions? Our Professional Service Experts are available for just this situation and many others. We will assist in the initial installation of Sawmill using best practices; work with you to integrate and configure Sawmill to generate reports in the shortest possible time. We will tailor Sawmill to your environment, create a customized solution, be sensitive to your requirements and stay focused on what your business needs are. We will show you areas of Sawmill you may not even be aware of, demonstrating these methods will provide you with many streamlined methods to get you the information more quickly. Often you'll find that Sawmill's deep analysis can even provide you with information you've been after but never knew how to reach, or possibly never realized was readily available in reports. Sawmill is an extremely powerful tool for your business, and most users only exercise a fraction of this power. That's where our experts really can make the difference. Our Sawmill experts have many years of experience with Sawmill and with a large cross section of devices and business sectors. Our promise is to very quickly come up with a cost effective solution that fits your business, and greatly expand your ROI with only a few hours of fee based Sawmill Professional Services. For more information, a quote, or to speak directly with a Professional services expert contact consulting@flowerfire.com.

Tips & Techniques: Using Cross-Reference Tables

Cross-reference tables (sometimes called cross-reference groups, or xrefs), are tables created by Sawmill in its back-end database. Cross-reference tables are generated during database builds and updates, and contain aggregated information from the main table of the database. For instance, a particular cross-reference table might contain one row for each day in a media server log dataset, with the number of accesses, play duration, bytes transferred, unique IPs, sessions, etc. for that day. When Sawmill generates an unfiltered Days report, it can generate it directly from this table, which is much smaller than the main table of the database; this allows it to generate this report, and other top-level reports, very quickly.

Cross-reference tables are created by default for each non-aggregating field in the database, which means there is roughly one cross-reference table for each report, so all top-level reports are boosted by cross-reference tables. Furthermore, every default cross-reference table also contains the date/time field, so any report can be filtered by date, and still use a cross-reference table. So all default unfiltered reports, and all default date-filtered reports, are accelerated using cross-reference tables.

If you add a new database field, however, or create a new report, or if you often use a particular combination of filters on a particular report, you may need to modify the cross-reference tables to ensure that this custom or filtered report is also fast. This newsletter gives an example of this.

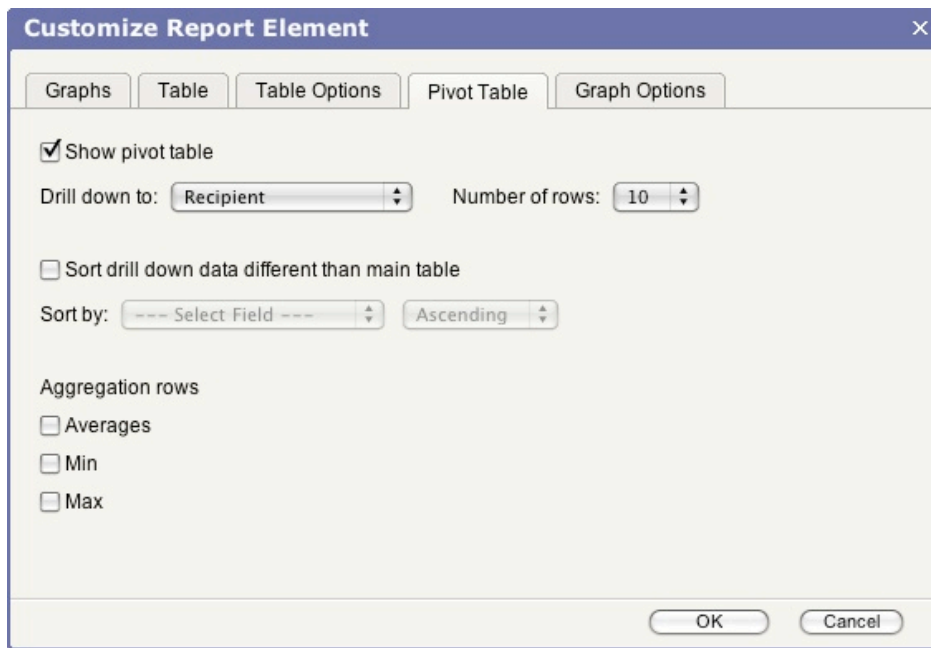
Building The Report

We'll start with a spam filtering server dataset. Clicking Senders shows the top sender domains in the data:

Senders					Customize Report in Config
01/Dec/2005 - 01/Jan/2006, 32 days (entire date range)					
Export Table Customize					
Item 1 - 10 of 237980		11 - 20 ▶		10 20 50 100 200 500 [...]	
Sender	Messages		Bytes		
1 (empty)	683,524	23.3 %	<div style="width: 23.3%;"></div>	3.63 G	
2 @hotmail.com	541,783	18.5 %	<div style="width: 18.5%;"></div>	933.62 M	
3 @yahoo.com	29,684	1.0 %	<div style="width: 1.0%;"></div>	799.56 M	
4 @ebay.com	14,670	0.5 %	<div style="width: 0.5%;"></div>	293.19 M	
5 @aol.com	13,298	0.5 %	<div style="width: 0.5%;"></div>	1.18 G	
6 @verizon.net	13,294	0.5 %	<div style="width: 0.5%;"></div>	82.32 M	
7 @returns.groups.yahoo.com	12,051	0.4 %	<div style="width: 0.4%;"></div>	66.59 M	
8 @endogenter.com	11,312	0.4 %	<div style="width: 0.4%;"></div>	26.79 M	
9 @denedia.freeimage.com	9,860	0.3 %	<div style="width: 0.3%;"></div>	30.05 M	
10 @cgwcorps.com	9,279	0.3 %	<div style="width: 0.3%;"></div>	23.12 M	
237970 other items	1,593,450	54.3 %	<div style="width: 54.3%;"></div>	17.61 G	
Total	2,932,205	100.0 %		24.62 G	

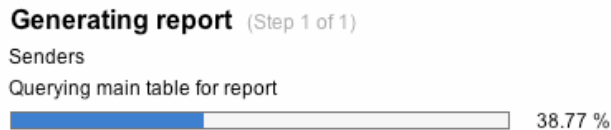
Senders Report

If we also want to know, in a single report, the top Recipients for each Sender, we can use a Pivot Table, by clicking Customize, then the Pivot Table tab, then "Show pivot table", and then selecting Recipients as the drill-down:



Customize Report Element: Add A Pivot

Clicking OK begins to generate this report. But it takes several minutes, and while it's thinking, it displays this progress bar:



Progress Of Report Using Main Table

The key phrase to watch for is "Querying main table." The main table is the primary table of the database, with one row for each, even in the log data. It can be millions or billions of lines long, depending on the dataset. Querying the main table can take many minutes, especially for an unfiltered report (where every single row must be aggregated to create the report), so it is to be avoided for any reports you use frequently. That's what xref tables are for, so let's cancel this report and head to Config to make it faster. In Config, click More Options, then Cross Reference Groups, and duplicate the existing Recipient group, calling it "Sender x Recipient", and add the Sender field to it:

The screenshot shows the Sawmill web interface. At the top, there is a navigation bar with the Sawmill logo and links for 'Config of profile spamfilter', 'View Reports', 'Admin', 'Logout (ferrari)', 'Help', and 'About'. Below this is a secondary navigation bar with links for 'Log Source', 'Log Filters', 'Log Processing', 'Database', 'Database Info', 'Reports', 'Report Options', 'Cross Reference Groups', and 'More Options'. A toolbar contains buttons for 'Save Changes', 'New Group', 'Duplicate', 'Delete', and 'Undo All Changes'. The main content area is divided into two panels. The left panel, titled 'Cross Reference Groups', shows a list of groups with checkboxes for various fields. The 'Sender x Recipient' group is selected and highlighted. The right panel shows the configuration for this group. It has a 'Name' field containing 'Sender x Recipient'. Below this are two lists: 'Active fields' and 'Available fields'. The 'Active fields' list contains 'Recipient', 'Sender', 'Date/time', 'Messages (Num)', and 'Bytes (Num)'. The 'Available fields' list contains 'Action', 'Attachment', 'Day of week', 'Geographic location', 'Hour of day', 'Keywords', 'Reason', 'Source hostname', 'Source IP', and 'Virus'. Between the lists are buttons for 'Remove >', 'Move up', and 'Move down'. At the bottom of the right panel, there is a checkbox labeled 'Use flat table'.

Adding A Cross-Reference Group

After adding a cross-reference group, we need to rebuild the database. After the build is complete, we can return to the reports, and again do the Recipients x Senders pivot table report. This time, the report comes up quickly:

Senders					Customize Report in Config
📅 01/Dec/2005 - 01/Jan/2006, 32 days (entire date range)					
Item 1 - 10 of 237980					
Export Table Customize					
Sender /	▼ Messages			Bytes	
1 (empty)					
1 (empty)	617,322	21.1 %	<div style="width: 21.1%;"></div>	3.24 G	
2 @mags.net	48,148	1.6 %	<div style="width: 1.6%;"></div>	198.42 M	
3 @hlla.com	6,906	0.2 %	<div style="width: 0.2%;"></div>	53.76 M	
4 @unisoft-cim.com	5,040	0.2 %	<div style="width: 0.2%;"></div>	32.54 M	
5 @taskmanagement.com	1,110	0.0 %	<div style="width: 0.0%;"></div>	13.31 M	
6 @tomortgageservices.com	432	0.0 %	<div style="width: 0.0%;"></div>	16.12 M	
7 @nutmegstamp.com	362	0.0 %	<div style="width: 0.0%;"></div>	2.97 M	
8 @2sbdigest.com	318	0.0 %	<div style="width: 0.0%;"></div>	2.75 M	
9 @ctweather.com	274	0.0 %	<div style="width: 0.0%;"></div>	1.10 M	
10 @westportcompany.com	185	0.0 %	<div style="width: 0.0%;"></div>	1.58 M	
142 other items	3,427	0.1 %	<div style="width: 0.1%;"></div>	80.82 M	
Sub total	683,524	23.3 %	<div style="width: 23.3%;"></div>	3.63 G	
2 @hotmail.com					
1 @hlla.com	452,714	15.4 %	<div style="width: 15.4%;"></div>	520.79 M	
2 @mags.net	50,171	1.7 %	<div style="width: 1.7%;"></div>	134.70 M	
3 @nutmegstamp.com	6,181	0.2 %	<div style="width: 0.2%;"></div>	39.49 M	
4 @risystems.com	3,988	0.1 %	<div style="width: 0.1%;"></div>	4.08 M	
5 @redmannguides.com	2,668	0.1 %	<div style="width: 0.1%;"></div>	2.63 M	

The Final Report

That's it--from now on, that report will be fast. As an added benefit, this also speeds up Recipient reports filtered on a particular Sender, or Sender reports filtered on a particular Recipient.

Which Fields Do You Need?

In order to get a particular report to use an xref group, you need all the fields in the report, plus all the fields in the filters. So if you have a report with three columns: Sender, Messages, and Bytes; and if you also have two filters on that report, a Date/time (date range) filter and a Recipient filter, you will need all those fields in the xref table: Sender, Recipient, Date/time, Messages, and Bytes.

Professional Services

This newsletter describes optimizing reports with cross-reference tables. This is one example of many types of optimization which are possible to make Sawmill build databases faster, and generate reports faster. Other possibilities include selective indexing, log splitting, field simplification, horizontal shrinkage (eliminating fields), vertical shrinkage (eliminating rows), and hierarchical cross-reference optimization. For large environments where performance is important, we recommend Sawmill Professional Services to help you quickly optimize your installation. If you need assistance with optimization, or with any other Sawmill tasks, our Sawmill Experts can help. Contact sales@sawmill.net for more information.

[Article revision v1.0]

[ClientID: 43726]