

“Providing Standardization to DHI (Dairy Herd Improvement) Databases with SAS and using SAS products to deliver user friendly applications.”

*Dr. Michael Tomaszewski, Texas A&M University, College Station, TX
Sandeep Gaudana, Texas A&M University, College Station, TX*

Abstract:

The ability to perform data mining is predicated upon access to similarly calculated variables. Dairy consultants and educators have attempted to mine dairy herd improvement (DHI) data but have been faced with lack of standardization among the several databases used by dairy producers. Using SAS Base, we have developed techniques to access base data from these different databases and calculate critical production indicators using similar algorithms. One of the databases is accessed by a scheduled SAS program that automatically obtains data from a FTP Server. All of this is done without any user intervention. SAS/Intrnet applications were developed to enable a user to request analysis via web. This paper discusses the steps taken to reach this stage of development and the techniques used to make programs efficient. Finally, “Management Matrix” – a SAS/Intrnet Application will be discussed.

Problem Definition:

When analyzing data, it is always better to have common variables calculated in a standardized manner. However, due to lack of calculation standardization of core parameters within the dairy industry, different organizations have developed central databases and software management programs that have not used standard calculation algorithms. It is this lack of standardization that faces the dairy consultants and educators who want to use these databases. SAS was used to overcome this problem.

Providing a one time standardized database requires heavy data cleaning and data manipulation techniques. After standardization a master database was developed which provides a single location for various decision support applications. When new data becomes available, they are added to the master database after necessary cleaning and manipulation. Several techniques have been used to perform this process. The design process of updating the database will be discussed. Finally the “Management Matrix” will be discussed which provides the users a web based analytical tool with the help of SAS/Intrnet.

Diagrammatic Representation:

Figure 1 shows an overview of how the process works. Databases accessed are DRMS, DairySTOR and DairyCOMP. The data is available independently. The Directory structure and the incoming data files are so named that they can be used effectively in the SAS programs. The importance of this is realized when using macros in the programs. As shown in the figure, the data from DRMS and DairySTOR are available from FTP websites. Programs have been developed that make this retrieval process automatic. Program code is discussed later. DairyCOMP data adds to the complexity as that data is not directly recognized by SAS. DairyCOMP software is used to create Excel files which are then used by SAS.

In the initial development stage, a SAS program was written to combine the existing data from all the sources. The resulting dataset was established as a master SAS database. In addition to this, a program was developed that adds any newly available data to this master database. This program runs whenever a new data file is available in the directory, thus updating the master database.

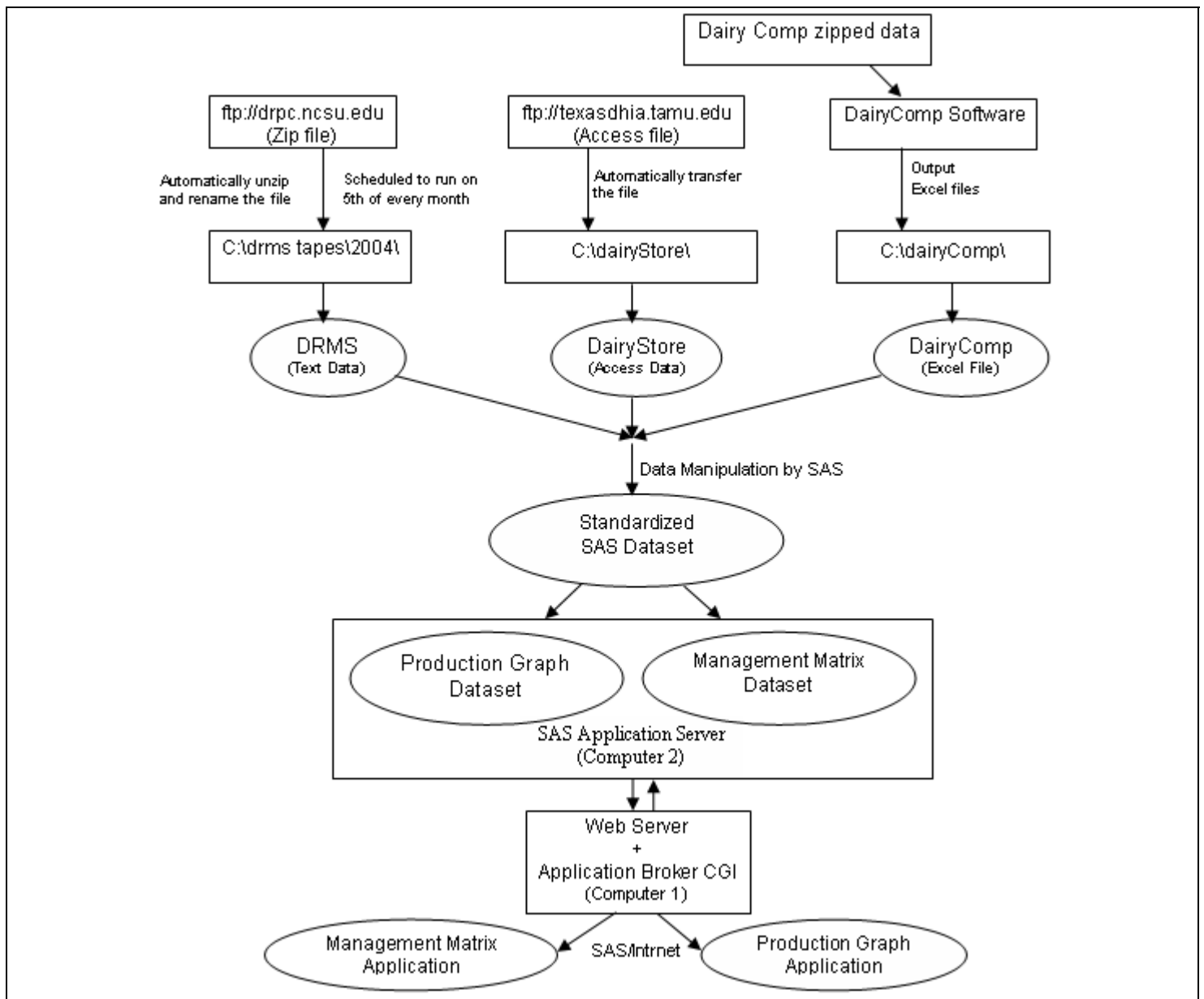


Figure 1: Diagrammatic Representation

Two web applications using SAS/Intrnet have been developed:

- a) Management Matrix
- b) Production Graph

In order to reduce the response time for a user request, sub-datasets for Production & Management Matrix data were created (Figure 1) which contain all the analysis for the desired variables. Whenever a user places a request for the analysis of certain variables the program uses these datasets to create the output.

Automating Data Retrieval Process:

Historically, DRMS zipped files were manually downloaded, unzipped and renamed in order to maintain a naming pattern. This was done every month.

For downloading the zip file, the FTP access method of the filename statement is now used to achieve automation. The X- command is used in order to unzip the files using pkunzip, rename the file, move the file to the required folder and thereafter to delete the zipped file to save system resources. Unfortunately, the directory name "DRMS TAPES", because it has a space in between the two names was a major problem when using the X-command. Since this directory was used in many other programs, renaming the directory would have required changes in these programs as well. Following is the code that finally worked:

```
/*Code for providing automation to download a zip file and storing the extracted file to
destination folder with required name*/

data _null_;

/*The data being obtained is always of the previous month i.e. in the month of July we
will get the data for June*/

    today = put (date()-31,worddate3.);
    month_d = put(date()-31,mddyy2.);
    year_d = put(date()-31,yymmdd2.);

/*The macro variable month stores the first three letters of the month*/
/*The macro variable m stores the month number */
/*The macro variable y stores the year number*/

    call symput('month',today);
    call symput('m',month_d);
    call symput('y',year_d);
run;

/*The required zip file is named as "texlact.jun" (example)*/

filename getftp FTP "texlact.&month"
    host="ftp.drpc.ncsu.edu"
    user="xxxxxxxx"
    pass="xxxx"
    cd="\fromdrms" lrecl=1 recfm=f;

filename putftp "c:\temp\drms&y&m..zip" recfm=f lrecl=1;

/*The following data step stores the zip file in c:\temp (Folder Temp should be
available) */

data _null_;
    infile getftp end=eof;
    input;
    file putftp;
    put _infile_;
run;
/*By using the option noxwait, X-commands does not wait for exit before returning to
SAS*/
/*The option xsync allows the SAS System and X- commands to execute synchronously*/

options noxwait xsync;
data _null_;
```

```

/*The following X-command unzips the file using pkunzip and again stores the extracted
file in c:\temp*/

x "call c:\pcdart\pkunzip.exe c:\temp\drms&y&m..zip c:\temp";

/*The following X-command renames the extracted file to our required format using macro
variables m and y*/

x "rename c:\temp\texlact.txt drms&y&m..txt";

/*The following statement finally worked by way of trial and error*/

x %unquote(%str('%') "move "c:\temp\drms&y&m..txt" "c:\drms tapes\20&y" " %str(%));

/*The following X-command deletes the downloaded zipped file from c:\temp*/

x "del c:\temp\drms&y&m..zip";

run;

```

DairySTOR was stored in a folder on the ftp server which was then imported in the destination folder for DairySTOR databases.

Automatically run SAS program when new data becomes available:

There was a desire to develop something that would lead to immediate update of the master database. The above code partially succeeded in getting the data automatically from the FTP website and transferring it to the required folder. But occasionally, the folder for the availability of new files needed to be tracked. One way was to schedule the running of WINDOWS© after a regular interval of time. However, instantaneous updating was desired. For this a small application called WATCHDIRECTORY was utilized. It's a small but very helpful application.

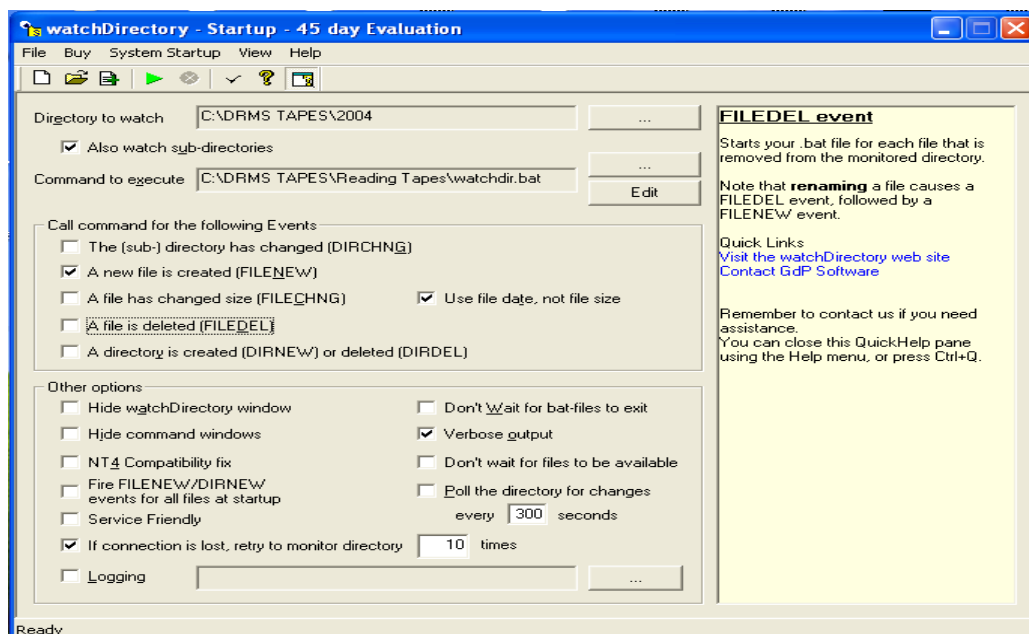


Figure 2: WATCHDIRECTORY Application to run SAS Program automatically

The .bat file that allows the required SAS program to run in batch mode can be created using notepad.exe and then can be saved as a .bat type . Following is the .bat file that was created.

```
"C:\Program Files\SAS\SAS System\9.0\SAS.exe" -sysin "c:\drms tapes\reading tapes\watchdir.sas"
```

Location of SAS.exe
Specifies the default location of SAS source code when run in Batch or interactive mode
Location of SAS Program to run

Using the WATCHDIRECTORY application, directory to be watched can be browsed and also the option to watch for any changes in sub directories can be selected.

There are different options available in the application to decide an event when .bat file has to be executed. The SAS program can be run in the batch mode when a new data file is created, deleted, renamed, overwritten or if there is a change in directory. Apart from this, there are several other options that can be used in this application.

Generating E-mail message in Data Step:

The program that updated the master database was run automatically when a new data file was available in the directory with the help of WATCHDIRECTORY application. However, the problems of run time errors were not solved. Thus a procedure was required that could help in regularly checking the log generated by the program. To make this task simpler, an e-mail message can be generated inside the data step. The e-mail message can contain a whole lot of information. An attachment of the log file along with the e-mail message could be sent. Using macros, information such as the number of records in the new data file and the name of the new file can be sent along with the email. The sample code for this program is as follows:

```
/*Assume that using FTP program explained earlier we have downloaded the file in the destination folder*/  
/*Also the WATCHDIRECTORY Application has detected the new file and has started running the program*/  
/*Suppose after some data manipulation on the data file we create a dataset "newcheck" which has to be added to Master Database*/
```

```
data _null_;  
    yr_d = put(date()-31,yymmdd2.);  
    mo_d = put(date()-31,mmdyy2.);  
    call symput('yr',yr_d);  
    call symput('mo',mo_d);  
run;
```

```
/*Using Microsoft Outlook as an Electronic Mail Program*/
```

```
filename mail EMAIL "sgaudana@tamu.edu"  
        Subject="SAS file drms&yr&mo..txt added to Master dataset";
```

```
/* "newcheck" is a data set created from the new data file "drms&yr&mo..txt" . The total number of records are stored in macro  
Variable "numrows"*/
```

```
proc sql noprint;  
    select count(*) into :numrows  
    from newcheck;  
quit;
```

```
data _null_;
```

```

file mail;
put "The new data file had &numrows records";

/*!EM_ATTACH! Sends an attachment along with the e-mail and !EM_SEND! Sends e-mail*/

PUT '!EM_ATTACH!("c:\drms tapes\2004\watchdir.log")';
PUT "!EM_SEND!";
run;

```

SAS/Intrnet:

After the base program was developed, SAS/AF was used to provide a better user interface. However, a procedure was needed that did not require the end user to have SAS software. Thus, SAS/Intrnet was used for this purpose.

Application Dispatcher component of SAS/Intrnet was used for this.

Architecture:

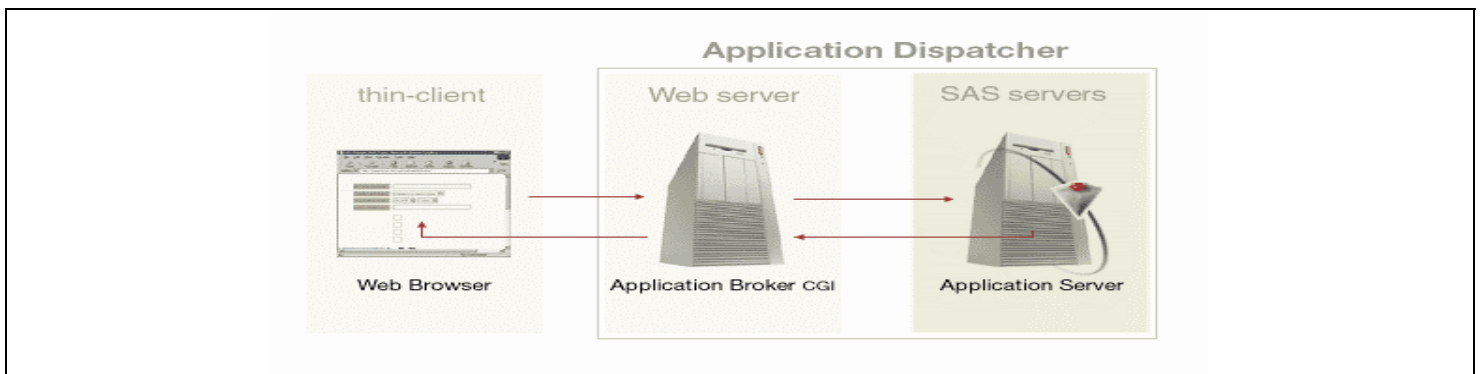


Figure 3: Architecture used for SAS/Intrnet Implementation

Figure 3 shows the basic architecture we have been using in order to provide the SAS/Intrnet Application. Presently, the number of Application Server is only one. But as traffic on the website increases it may be necessary to use multiple Application Servers. Thereafter, Application Load Manager can be utilized in order to optimize the resources of Application Dispatcher on the network.

The following is an introduction to Management Matrix which is one of the applications provided on the web. Dreamweaver© software was used to develop the web application.

Management Matrix Application:

The Management Matrix allows the user to form his own contemporary groups, called peergroup. A peergroup is a group of herd codes (Each herd code refers to a Dairy Farm) which comprise a cohort that has some similar characteristics. There is also an option for “No Peergroup”. By selecting this option, the user receives an analysis of a single herd code.

DHI SOUTHWEST
Dairy Herd Improvement
TEXAS A&M UNIVERSITY, COLLEGE STATION

Select PeerGroup Refresh

Peer Group			
No Peergroup	Select		
Sample	Select	Edit	Delete
SouthWest	Select	Edit	Delete

Enter New Peer Group

BACK

Figure 4: Management Matrix Application

This application further provides features for creating a new peergroup, editing an existing peergroup and deleting the entire peergroup. The description of these features is as explained below:

1) Enter New PeerGroup:

As shown below in Figure 5, clicking on “Enter new Peergroup” on the main page shown in Figure 4 takes the user to webpage 1 where one can enter a new Peergroup which then directs him to webpage 2 for addition of herd codes to this new Peergroup. After completion the user is taken back to the main page which has now been populated with a new Peergroup.

DHI SOUTHWEST
Dairy Herd Improvement
TEXAS A&M UNIVERSITY, COLLEGE STATION

Enter name of new PeerGroup Insert

1

DHI SOUTHWEST
Dairy Herd Improvement
TEXAS A&M UNIVERSITY, COLLEGE STATION

Enter name of new PeerGroup Insert

Add Herdcodes

Peergroup

Herdcode Add

Present Herdcodes

Main Menu

2

DHI SOUTHWEST
Dairy Herd Improvement
TEXAS A&M UNIVERSITY, COLLEGE STATION

Select PeerGroup Refresh

Peer Group			
No Peergroup	Select		
NORTH WEST	Select	Edit	Delete
Sample	Select	Edit	Delete
SouthWest	Select	Edit	Delete

Enter New Peer Group

BACK

Figure 5: Creating a New Peergroup

2) EDIT:

Here the user can click on the Edit Button corresponding to the Peergroup to be edited. In this way, a new herdcode may be added or deleted from the peergroup. Click on Edit Button corresponding to Peergroup “Sample” shown in Figure 5.

On clicking the edit Button, a Pop-Up Window opens up. Here, the user can select to add Herdcode from the available list. Also, the user has an option of selecting multiple Herdcodes to delete at a time. To close the window and return to the main page the user can click on “Close Window” button.

Edit Peergroups - Microsoft Internet Explorer

Peergroup : Sample

Add Herdcode Insert

74870032
74880009
74880013
74940023
74997002

Select Herdcode to Delete

☐ 74158787
☐ 74450435
☐ 74460437
☐ 74460451

Delete Close Window

Figure 6: Edit Peergroup

As shown in Figure 6 there is a list of herdcodes to choose from. This list is dynamic. That means, whenever the master database is updated, this list should also be updated to accommodate new herdcodes or delete the unused herdcodes.

SQL Server is used as a back end for this application. Following libname statement is used in SAS to establish a connection with SQL Server with the help of SAS/Access.

```
Libname sqldata odbc dsn=xxxx ;
```

Here, “sqldata” is a libref. With the help of odbc, SAS can work on the tables in SQL Server as if they were SAS datasets. It can also create new tables or delete existing tables provided one has got sufficient authorization on the tables.

3) Delete:

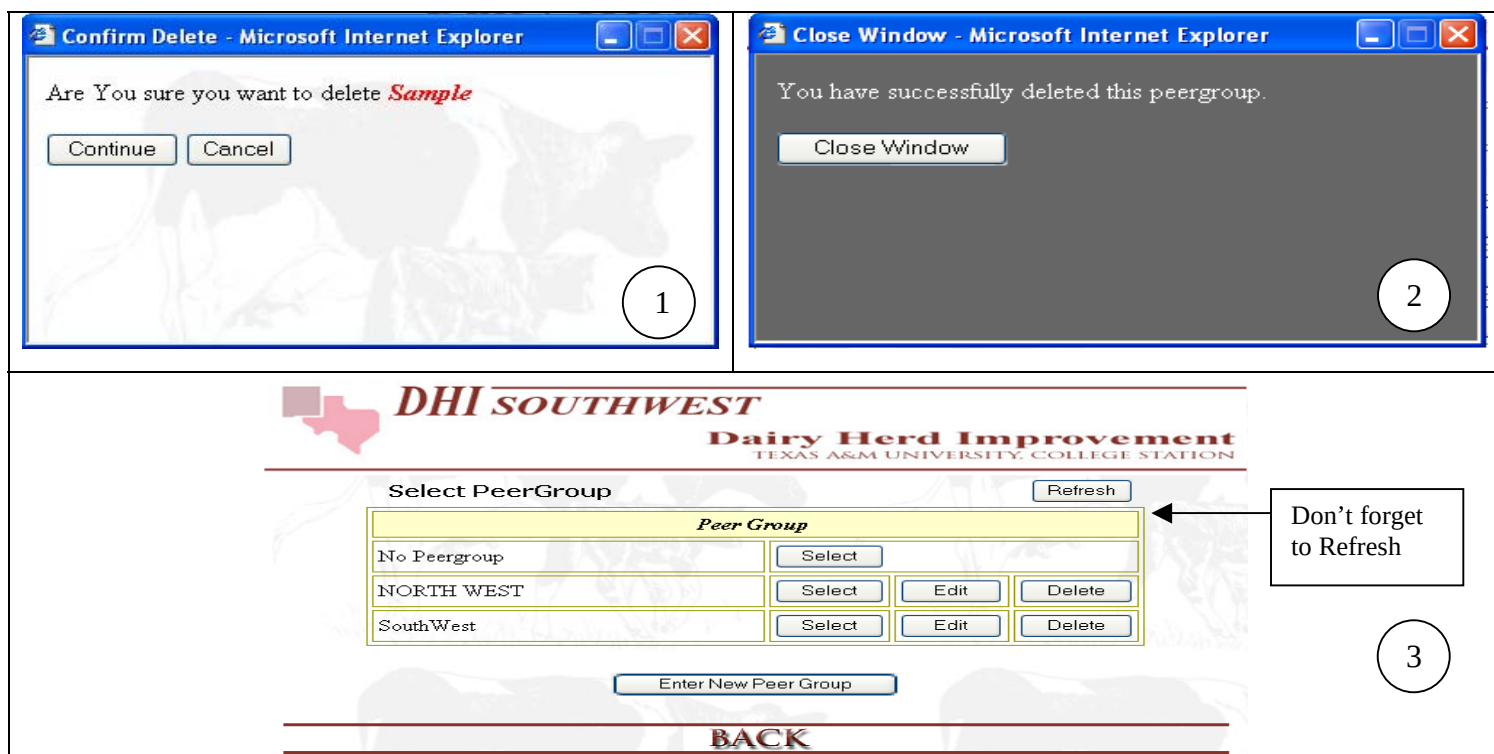


Figure 7: Delete Peergroup

4) Select:

Clicking on the Select Button on the main page corresponding to the peergroup will take the user to a webpage wherein the user can select different variables he wants to analyze and show up in the output.

The application in Figure 8 requires a bit of JavaScript coding but provides a user friendly application. Management Matrix allows users to have analysis for the following details:

- a) Herd Structure
- b) Production
- c) Reproduction

- d) Udder Heath
- e) Culling



DHI SOUTHWEST

Dairy Herd Improvement
TEXAS A&M UNIVERSITY, COLLEGE STATION

You have selected Peer Group : **NORTH WEST**

Herd Structure Select All Number Of Cows % Cows In Milk % Cows In Milk (First) % Cows In Milk (Second) % Cows In Milk (Third) Cows Entered Cows Left << Remove Production Select All TD Milk (ALL) TD Milk (Cows In Milk) First Milk (1st Lact) First Milk Avg DIM (1st) First Milk (2nd Lact) First Milk Avg DIM (2nd) First Milk (3rd Lact) First Milk Avg DIM (3rd) Peak Milk (1st Lact) Peak Milk (2nd Lact) Peak Milk (3rd Lact) Peak Milk (All) Summit Milk (1st Lact) Summit Milk (2nd Lact) Summit Milk (3rd Lact) Summit Milk All << Remove Reproduction Select All Calving Interval Projected Calving Interval Days To First Service Days Open Days Open (P) Services/Preg (P) Conception Rate Cows Pregnant (<150 DIM) Cows Pregnant (>150 DIM) % Successful Breeding (1st) % Successful Breeding (2nd) % Successful Breeding (3rd) Days Dry << Remove Udder Health Select All Somatic Cell (0 - 40) Somatic Cell (41 - 100) Somatic Cell (101 - 200) Somatic Cell (201 - 300) Somatic Cell (> 300) << Remove Culling Select All Cows Died (Prev Month) Cows Died (12 Months) Total Number Left (Month) Annualized Left Cows Leaving (1 - 40) Cows Leaving (41 - 100) Cows Leaving (101 - 200) Cows Leaving (201 - 300) Cows Leaving (> 300) << Remove	Herd Structure - Selected Production - Selected Reproduction Selected Udder Health - Selected Culling - Selected
---	--

● HTML ○ PDF

Continue
Cancel

Figure 8: Selection of Variables for Analysis

Again there are variables in each of the details. This web application allows the user to select all the variables together by clicking on “Select All” or individually select the variable by clicking on the variable name. Again the user can remove the variables already selected from the right side list by selecting them and then clicking on “<< Remove” Button.

Finally, the user can have either HTML or PDF report. ODS feature of SAS makes the job much simpler.

HTML Output:

Following is the sample HTML Output obtained by selecting “NORTH WEST” peergroup.

	<table><tr><th colspan="6">Detail= Herd Structure</th></tr><tr><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><th>Type</th><th>74460437 12/26/03</th><th>74460451 01/23/04</th><th>Min</th><th>Max</th><th>Ave</th></tr><tr><td>Number Of Cows</td><td>907</td><td>1616</td><td>907</td><td>1616</td><td>1262</td></tr><tr><td>% Cows In Milk</td><td>69%</td><td>78%</td><td>69%</td><td>78%</td><td>74%</td></tr><tr><td>Cows Entered</td><td>15</td><td>51</td><td>15</td><td>51</td><td>33</td></tr><tr><td>Cows Left</td><td>25</td><td>41</td><td>25</td><td>41</td><td>33</td></tr></table>	Detail= Herd Structure												Type	74460437 12/26/03	74460451 01/23/04	Min	Max	Ave	Number Of Cows	907	1616	907	1616	1262	% Cows In Milk	69%	78%	69%	78%	74%	Cows Entered	15	51	15	51	33	Cows Left	25	41	25	41	33	
Detail= Herd Structure																																												
Type	74460437 12/26/03	74460451 01/23/04	Min	Max	Ave																																							
Number Of Cows	907	1616	907	1616	1262																																							
% Cows In Milk	69%	78%	69%	78%	74%																																							
Cows Entered	15	51	15	51	33																																							
Cows Left	25	41	25	41	33																																							
	<table><tr><th colspan="6">Detail= Production</th></tr><tr><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><th>Type</th><th>74460437 12/26/03</th><th>74460451 01/23/04</th><th>Min</th><th>Max</th><th>Ave</th></tr><tr><td>TD Milk (All)</td><td>52.7</td><td>59.9</td><td>52.7</td><td>59.9</td><td>56.3</td></tr><tr><td>TD MILK (Cows In Milk)</td><td>76.2</td><td>76.6</td><td>76.2</td><td>76.6</td><td>76.4</td></tr><tr><td>Peak Milk (All)</td><td>97.7</td><td>92.5</td><td>92.5</td><td>97.7</td><td>95.1</td></tr><tr><td>Summit Milk (All)</td><td>91.8</td><td>84.5</td><td>84.5</td><td>91.8</td><td>88.2</td></tr></table>	Detail= Production												Type	74460437 12/26/03	74460451 01/23/04	Min	Max	Ave	TD Milk (All)	52.7	59.9	52.7	59.9	56.3	TD MILK (Cows In Milk)	76.2	76.6	76.2	76.6	76.4	Peak Milk (All)	97.7	92.5	92.5	97.7	95.1	Summit Milk (All)	91.8	84.5	84.5	91.8	88.2	
Detail= Production																																												
Type	74460437 12/26/03	74460451 01/23/04	Min	Max	Ave																																							
TD Milk (All)	52.7	59.9	52.7	59.9	56.3																																							
TD MILK (Cows In Milk)	76.2	76.6	76.2	76.6	76.4																																							
Peak Milk (All)	97.7	92.5	92.5	97.7	95.1																																							
Summit Milk (All)	91.8	84.5	84.5	91.8	88.2																																							
	<table><tr><th colspan="6">Detail= Reproduction</th></tr><tr><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><th>Type</th><th>74460437 12/26/03</th><th>74460451 01/23/04</th><th>Min</th><th>Max</th><th>Ave</th></tr><tr><td>Calving Interval</td><td>15.7</td><td>14.4</td><td>14.4</td><td>15.7</td><td>15.0</td></tr><tr><td>Projected Calving Int.</td><td>17.8</td><td>13.8</td><td>13.8</td><td>17.8</td><td>15.8</td></tr><tr><td>Days To First Service</td><td>172.9</td><td>87.2</td><td>87.2</td><td>172.9</td><td>130.0</td></tr><tr><td>Days Dry</td><td>82.4</td><td>97.2</td><td>82.4</td><td>97.2</td><td>89.8</td></tr></table>	Detail= Reproduction												Type	74460437 12/26/03	74460451 01/23/04	Min	Max	Ave	Calving Interval	15.7	14.4	14.4	15.7	15.0	Projected Calving Int.	17.8	13.8	13.8	17.8	15.8	Days To First Service	172.9	87.2	87.2	172.9	130.0	Days Dry	82.4	97.2	82.4	97.2	89.8	
Detail= Reproduction																																												
Type	74460437 12/26/03	74460451 01/23/04	Min	Max	Ave																																							
Calving Interval	15.7	14.4	14.4	15.7	15.0																																							
Projected Calving Int.	17.8	13.8	13.8	17.8	15.8																																							
Days To First Service	172.9	87.2	87.2	172.9	130.0																																							
Days Dry	82.4	97.2	82.4	97.2	89.8																																							
	<table><tr><th colspan="6">Detail= Culling</th></tr><tr><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><th>Type</th><th>74460437 12/26/03</th><th>74460451 01/23/04</th><th>Min</th><th>Max</th><th>Ave</th></tr><tr><td>Cows Died (Prev Month)</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>Cows Died (12 Months)</td><td>5</td><td>61</td><td>5</td><td>61</td><td>33</td></tr><tr><td>Total Number Left (Month)</td><td>16</td><td>0</td><td>0</td><td>16</td><td>8</td></tr><tr><td>Annualized Left</td><td>202</td><td>416</td><td>202</td><td>416</td><td>309</td></tr></table>	Detail= Culling												Type	74460437 12/26/03	74460451 01/23/04	Min	Max	Ave	Cows Died (Prev Month)	0	0	0	0	0	Cows Died (12 Months)	5	61	5	61	33	Total Number Left (Month)	16	0	0	16	8	Annualized Left	202	416	202	416	309	
Detail= Culling																																												
Type	74460437 12/26/03	74460451 01/23/04	Min	Max	Ave																																							
Cows Died (Prev Month)	0	0	0	0	0																																							
Cows Died (12 Months)	5	61	5	61	33																																							
Total Number Left (Month)	16	0	0	16	8																																							
Annualized Left	202	416	202	416	309																																							

Figure 9: HTML Output for Management Matrix

When it comes to web applications, everyone is concerned with the response time required for output. Following points should be considered when providing output with SAS/Intrnet:

- 1) Only keep the essential details in the dataset that is being accessed. For example, there may be some herdcodes which are no longer accessed. Remove them whenever the dataset is updated. Use length statement to reduce overall size of the database. This will allow more observations to fit in a single data set memory page which can reduce the number of I/O (Input/Output) operations to read a SAS data set.
- 2) Our application has a master database which is 1 GB in size. With a SAS program all the calculations for the desired variables can be made and then the output can be provided. But instead of accessing the entire database, whenever the master database is updated, a program that will make calculations for all the herdcodes with all the necessary variables is run. The resulting dataset is of 145 KB.
- 3) In case of heavy traffic, try to provide a dedicated computer for SAS Application Server.
- 4) In case of multiple Application Servers one can use Application Load Manager to decide the usage of a particular Application Server.

Conclusion:

Standardization to the DHI databases is provided by data manipulation techniques in SAS. Also, a certain level of automation was provided to the update process of the master database. Using SAS/Intrnet and some other vendor products user friendly web applications can be provided.

References:

SAS Sample Programs at:

http://ftp.sas.com/techsup/download/sample/samp_lib/hostsampBase_SAS_Sample_Programs.html

SAS Documentation at:

<http://www2.sas.com/proceedings/sugi29/052-29.pdf>

Contact Information:

Contact the authors at:

Dr. Michael Tomaszewski
Dairy Informatics Laboratory
2471 TAMU
Texas A&M University
College Station, TX-77843
E-mail – mat@dairy.tamu.edu

Sandeep Gaudana
Texas A&M University
401, Stasney Street,
Apartment #504,
College Station, TX-77840
E-mail – sgaudana@tamu.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.