Quick and Dirty Survey Analysis to Assess Real-Time Requirements
Author: Joseph Caruso PhD, IBM Corporation, Poughkeepsie, NY
Co-Author: Salvatore Potenza PMP, IBM Corporation, Poughkeepsie, NY

ABSTRACT

New web-based applications are developed every day in large corporations. Since the functionality required of these web-based applications far exceeds the standards of older legacy applications, there is often a trade-off between performance and function. To minimize the effect on performance, developers have recently been requiring that read-only users who do not require update capability go to an information warehouse (IW) to satisfy their data needs.

This often creates conflict between end users and developers. End users who are familiar with real-time access to data are reluctant to accept data from an IW that is updated nightly. This paper describes the methodology used to assign users between a web-based system and the IW. A list of simple questions was developed and sent to users of the legacy system. This process was implemented using SAS®.

SAS® code was written to randomize the legacy system active user list and generate e-mail addresses for 300 users. A fifty percent response rate was received within two days and results were tabulated using Excel®. A recommendation on how to split the user base between the web-based system and the IW was presented to management on short notice.

INTRODUCTION

Management decisions in industry are often made with an incomplete set of data and facts surrounding the issues being addressed. Possible reasons for the incomplete data are the need to make decisions in a timely fashion, the inability to measure the information, the cost of obtaining the information, or the business knowledge or skills available to obtain the information.

When these decisions involve groups of people survey analysis can be a useful source of information to supply the missing facts. However, many analysts and executives do not understand the benefits of using a survey to help support decision analysis. In addition, survey analysis is usually considered an expensive and untimely option and is therefore ruled out. This paper brings survey analysis back as an option in the day-to-day management decision making process by minimizing the resource required and maximizing the timeliness of data collection to support the decision analysis.

The rest of this paper describes a “quick and dirty” survey analysis to assist with a management decision where a new web based software tool was replacing a legacy “green screen” tool (CMCAR) within the United States. The web based system would have more functionality but there were concerns about the performance of the new tool. The issue to be addressed was how many of the 1,400 active users of the legacy tool would have access to the new real-time web based tool and how many would be routed to an information warehouse with identical information but which was only updated nightly.

METHODOLOGY

From analysis of the existing green screen legacy tool it was determined that there were 1,400 active users over the past 3 months. The users were split across a cross functional organization into 650 CSO (Customer Service Officers) and 750 non-CSO users. The CSO users were thought to be more likely to require real-time use of the system. What needed to be determined was how many of each of these user groups needed real-time access.
Also, some auxiliary information would be needed such as whether the users were familiar with an information warehouse, why they needed real-time access, what data fields are queried, etc. to address follow-on issues.

The high level methodology was to:
1. Develop a simple yes/no questionnaire that could be put in an e-mail
2. Write SAS® code to convert the active user list to a random sample of e-mail addresses for each group (CSO and non-CSO)
3. Send an e-mail to the randomized list for each group
4. Key the results into a simple 0/1 Excel® spreadsheet
5. Analyze the results using an Excel® pivot table
6. Build a presentation with recommendations to management

THE QUESTIONNAIRE

The questionnaire limited most responses to “check all that apply” and “Yes/No” type responses that could easily be recorded. There were only 2 open ended questions which required text entry (which were copied into the Excel® spreadsheet on return). The open ended questions were “other (specify)” and the last question, “How would not having real-time affect your job?”

1) What do you primarily use CMCAR for? (Check all that apply)
   a) search for base agreements _______
   b) other (specify) _______

2) What data do you use to search for base agreements in CMCAR? (check all that apply)
   a) customer number __________
   b) enterprise number __________
   c) customer number/ form number ________
   d) enterprise number/ form number ________

3) Have you ever used an information warehouse (IW) to access information? (y/n) ________ (examples would be MMS, FIW, Edge on Demand)
   Have you heard of Edge on Demand? (y/n) ________
   If yes, have you used Edge on Demand? (y/n) ________

4) Would having agreement data up updated daily (rather than real-time or instantaneously) have any adverse impact on your job responsibilities? (y/n) ________
   How would not having real-time data affect your job? ____________________________________

SAS® CODE TO CREATE A RANDOMIZED LIST

A list of the active users for the CSO and non-CSO groups was obtained from the developers for the existing legacy application. This user list included the name and the short user e-mail address. SAS® code was written to randomly sort the list and then select the first 150 users from each group out of the total 1400 active users as targets for the questionnaire.

Based on an initial trial of 50 users it was determined that a 50% response rate could be expected. This would result in 75 responses for each group. This was determined to be a reasonable sample size for this survey where 80% accuracy might be a more reasonable standard than the typical 95-99% accuracy required when there is a higher cost for a wrong decision.

The RANUNI function was used to randomize the sort order of the list and the first 150 users were selected for each group. The SEED input (12580) to the RANUNI function is a starting point for the algorithm and should be
randomly selected with at least 5 digits. Randomizing the survey sample is a key element in eliminating unwanted bias in results.

"READ IN ACTIVE USER LIST"
PROC IMPORT OUT= WORK.master
   DATAFILE= "C:\Documents and Settings\Administrator\My Documents\carusodata\OPPT\noncso user list 030807.xls"
   DBMS=EXCEL®2000 REPLACE;
   GETNAMES=YES;
RUN;

/*READ IN e-mail IDs and concatenate corporate address*/
data master;
   set master;
   aasid=scan(aasid_node_e-mail,1);node=scan(aasid_node_e-mail,2);e-mail=scan(aasid_node_e-mail,3);
   lname=scan(aasid_node_e-mail,4);fname=scan(aasid_node_e-mail,5);
   length e-mailid 30.;e-mailid=trim(e-mail)||'@us.ibm.com,';
run;

/*KEEP US USERS ONLY*/
data redmaster;
   retain e-mailid obs group aasid node e-mail lname fname z;

/*CREATE UNIFORM RANDOM NUMBER Z*/
set master;
   keep e-mailid obs group aasid node e-mail lname fname z;
   z=ranuni(12580);
   if node='IBMUS';
run;

/*SORT LIST BY RANDOM NUMBER*/
proc freq data=redmaster; tables group; run;
proc sort data=redmaster; by z; run;

/*KEEP FIRST 150 USERS*/
data sample; set redmaster; if _n_<=150;run;
proc freq data=sample; tables group; run;

/*OUTPUT RANDOM SAMPLE OF E-MAILS TO SPREADSHEET*/
PROC EXPORT DATA= WORK.REDMASTER;
   OUTFILE= "C:\Documents and Settings\Administrator\My Documents\carusodata\OPPT\noncso user list updated.xls"
   DBMS=EXCEL®2000 REPLACE;
RUN;

SEND QUESTIONS TO USERS
The randomized target e-mail addresses that were printed to the Excel® spreadsheet in the previous step were cut and pasted to the “send to” address in an e-mail which contained the questionnaire. The e-mail had instructions for the respondents to edit the e-mail with their responses and send them back to the sender. The e-mails were then submitted.
DATA COLLECTION

Responses were received from 142 users within 2 days. This is a much higher response rate (about 50%) that one might expect from an external “business-to-business” survey where responses are as low as 20%. The reason for the higher response rate was attributed to the direct impact the new system would have on the users of the system. The users were anxious to give their answers to the questionnaire so that proper consideration would be given to their needs.

As the results came back from users the e-mails were received and the randomized spreadsheet used to generate the e-mails was expanded with columns for 0/1 (no/yes) input. Each e-mail was opened and results were manually recorded on the matching e-mail record in the spreadsheet. The text responses (open ended) were cut and pasted in the rightmost columns.

ANALYSIS

Raw data analysis of the responses received was done using Excel® pivot tables. Only CSO respondents actually use the legacy tool to for updating information. The raw data below clearly shows that the non-CSO category has more of a propensity to use a data warehouse and less of a propensity to require real-time analysis. Auxiliary data not pertinent to the analysis of the real-time vs data warehouse decision is not shown here.

<table>
<thead>
<tr>
<th>CMCAR Respondents</th>
<th>CSO</th>
<th>NONCSO</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Update Capability</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Real-time Required-No Warehouse Experience</td>
<td>8</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>Real-time Required - Warehouse Experience</td>
<td>14</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>No Real-time Required-No Warehouse Experience</td>
<td>21</td>
<td>33</td>
<td>54</td>
</tr>
<tr>
<td>No Real-time Required - Warehouse Experience</td>
<td>17</td>
<td>26</td>
<td>43</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>72</td>
<td>142</td>
</tr>
</tbody>
</table>

The following is a graphical analysis of the above data. The real-time propensity moves from highest to lowest as you move from left to right. The real-time required colors (blue, red, yellow) are 46% for CSO and only 18% for non-CSO users.
PRESENTATION TO MANAGEMENT

In order to reflect the actual population for a management decision the sample data needed to be weighted to the full US population of users (the original 1400) and with an additional 15% increase for the fact that the new web based solution would be available to Canada as well. The following table was presented to management.

<table>
<thead>
<tr>
<th>Extrapolation to Total Population (1618)</th>
<th>CSO</th>
<th>NONCSO</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Update Capability</td>
<td>108</td>
<td>0</td>
<td>108</td>
</tr>
<tr>
<td>Real-time Required-No Warehouse Experience</td>
<td>86</td>
<td>84</td>
<td>170</td>
</tr>
<tr>
<td>Real-time Required - Warehouse Experience</td>
<td>151</td>
<td>72</td>
<td>223</td>
</tr>
<tr>
<td>No Real-time Required-No Warehouse Experience</td>
<td>227</td>
<td>395</td>
<td>622</td>
</tr>
<tr>
<td>No Real-time Required - Warehouse Experience</td>
<td>183</td>
<td>311</td>
<td>495</td>
</tr>
<tr>
<td>Total</td>
<td>756</td>
<td>863</td>
<td>1618</td>
</tr>
</tbody>
</table>

From the table above 5 alternatives were derived to present to management. An upper limit, lower limit, expected users, and dissatisfied users was derived for each alternative. The lower limit was taken to be the number of users that require real-time of those that are allowed to have real-time. The upper limit is the number of users that are allowed to have real time. Expected users is the average of the upper and lower limits. Dissatisfied users is the number of users who require real-time access who would not be granted real-time access.

Alternative 2 (grant real-time access to CSO users and screen non-CSO users) was recommended and accepted by management as the appropriate alternative to implement. This alternative was a balance between letting all users have real-time access with high impact to performance (alternative 1) and high levels of screening (alternative 4) while at the same time eliminating dissatisfied users (alternatives 3 and 5). Also, if users learn to use the data warehouse over time then the number of users could approach the lower limit.

<table>
<thead>
<tr>
<th>Alternative</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>Expected Users</th>
<th>Potential Dissat Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 All users have option of real time access</td>
<td>501</td>
<td>1618</td>
<td>1060</td>
<td>0</td>
</tr>
<tr>
<td>2 CSO Users have real time access, Non CSO have real time access by special request</td>
<td>501</td>
<td>511</td>
<td>706</td>
<td>0</td>
</tr>
<tr>
<td>3 CSO Users Have Real Time access</td>
<td>345</td>
<td>756</td>
<td>550</td>
<td>156</td>
</tr>
<tr>
<td>4 Only those with Real Time Need</td>
<td>501</td>
<td>501</td>
<td>501</td>
<td>0</td>
</tr>
<tr>
<td>5 Only Users Requiring Update Capability have real time access</td>
<td>108</td>
<td>108</td>
<td>108</td>
<td>383</td>
</tr>
</tbody>
</table>
CONCLUSIONS

The entire process for conducting the “quick-and-dirty” survey took one week from questionnaire design to management decision. The cost was one part time person for the week. The information gathered quantified and quelled a debate between development and the users that had no signs of going away. In addition there were other benefits of the survey. Some examples of unexpected feedback follow:

This is great info! Is there a way to identify the name of the person who made the first comment in the "Other uses..." section?

I have a team that is heavily dependent on CMCAR and are trying to develop tools that link with CMCAR so we are very interested in the effort and are willing to assist in any way that we can.

In summary, a quick-and-dirty survey can provide valuable information to management with minimal cost and on quick notice. The survey may also have unintended exploratory and awareness benefits as seen above. Randomizing your list of target respondents to minimize bias is a key component of the analysis and gives the survey additional credibility. Extrapolation to the full population (weighting) and analysis by groups of interest (stratification) also adds to survey credibility. If the survey is kept simple and combined with business knowledge one can utilize SAS® capabilities to bring necessary information to the decision making process.

REFERENCES


CONTACT INFORMATION

Joseph Caruso                        Salvatore Potenza
IBM Corporation /P184                IBM Corporation / P184
2455 South Road                     2455 South Road
Poughkeepsie NY, 12601-5400         Poughkeepsie, NY 12601-5400
jcaruso@us.ibm.com                  spotenza@us.ibm.com