THE STATISTICAL ANALYSIS OF GLOBAL GROWTH (SAGG) PROJECT

John C. Kupecki, Senior Computer Specialist
Veterans Affairs Information Systems Center
Albany, New York

Kornel A. Krechoweckyj, Senior Computer Specialist
Veterans Affairs Information Systems Center
Newington, Connecticut

ABSTRACT

The Department of Veterans Affairs (VA) has been running a hospital information system under the name of Decentralized Hospital Computer Program (DHCP) at its medical centers since 1984. The expansion of this program led to growing field concerns about the unchecked growth of the DHCP databases and the lack of additional tools to manage that growth. The Veterans Health Administration (VHA) began development of the Statistical Analysis of Global Growth (SAGG) Project in early 1992. The project monitors and tracks DHCP database activity at sites through regular global statistical data collection and transmission into a single VHA-wide MUMPS database. Key components are a fully automated design, minimal impact on hospital resources, and the immediate feedback of reliable information. The software collects information similar to the MUMPS global efficiency routines and is compatible with the current M database systems at the medical centers. The SAGG Project produces statistical reports for the medical centers and other organizations within the VA in order to better manage disk resources.

DISCUSSION

Over the years, there has been a tremendous growth in the size and scope of the Decentralized Hospital Computer Program (DHCP) at the medical centers within the Department of Veterans Affairs (VA). This rapid expansion has included increases both in hardware configurations and software packages. Presently, there are over fifty nationally released DHCP software packages, each contributing to the expanding databases at VA Medical Centers (VAMCs). This expansion has increased the complexity of system management adding additional strain on both computer and human resources.

The VA realized that reliable quantitative information had to be obtained in order to accurately track global database growth patterns at the medical centers. These statistics would allow the DHCP administrators to more efficiently address the growing field concerns among the various VAMCs. Such a database would help in the recognition and correction of any emerging problems prior to their causing operational difficulties. Reliable information would allow validation of current sizing model algorithm predictions and possibly point to new ways of interpretation.

In response to this need, the Veterans Health Administration (VHA) developed the Statistical Analysis of Global Growth (SAGG) Project in early 1992. Developed primarily as a statistical tool that examines global database sizes and efficiencies, SAGG incorporates other key features into the project. This fully automated MUMPS package regularly monitors DHCP global activity at each site with only minimal impact on the computer center's resources. The software is compatible with all current M database systems running at the sites and is easy to manage. Also, pertinent information relating to the captured data is immediately transmitted back to the participating site. Lastly, the captured information merges into a centralized SAGG database that utilizes the VA developed FileMan database management system. Subsequently, a variety of statistical analyses are performed and formulated into different reports.

FUNCTIONAL DESCRIPTION

The SAGG Project software fully utilizes the capabilities of the VA developed Kernel modules that currently run at the VA Medical Centers. The SAGG routines are based on three Kernel modules: TaskMan, MailMan and FileMan. TaskMan provides the scheduling interface and MailMan supplies the software interface to both local and network electronic mail systems. FileMan stores the global information from the individual sites in the centralized SAGG database. Collection runs are scheduled on a monthly basis and the captured data is merged into the SAGG database to permit trending and comparisons against previously captured information.

Installation of the SAGG Project software at a participating site creates the necessary components of the package. Several events occur during package initialization including the placement of the SAGG data collection routines and the creation of a local data file. This local file contains the placement information for the temporary collection global (^A1B5GE) which will hold the actual global efficiency data for the designated production volume sets. Additionally, this file stores the names of all production volume sets on the system that will be analyzed. Also, the initialization phase creates a local SAGG mail group which will receive all SAGG Project notification messages.

The package uses the VA developed TaskMan utility to schedule the initial global collection cycle and reschedules itself monthly. The fully automated data collection cycle captures global efficiency information into a temporary collection global (^A1B5GE). The SAGG data collection routines gather information similar to the MUMPS global efficiency routines. However, rather than displaying or printing this information to a terminal or printer, the SAGG routines differ significantly by directly storing the global efficiency data into the ^A1B5GE temporary data collection global. Besides obtaining the number of global pointer and data blocks and their global efficiency levels, the SAGG data collection routines also obtain: site name, operating system type and version, M database system type and version, and information that is pertinent to VA developed files and packages.

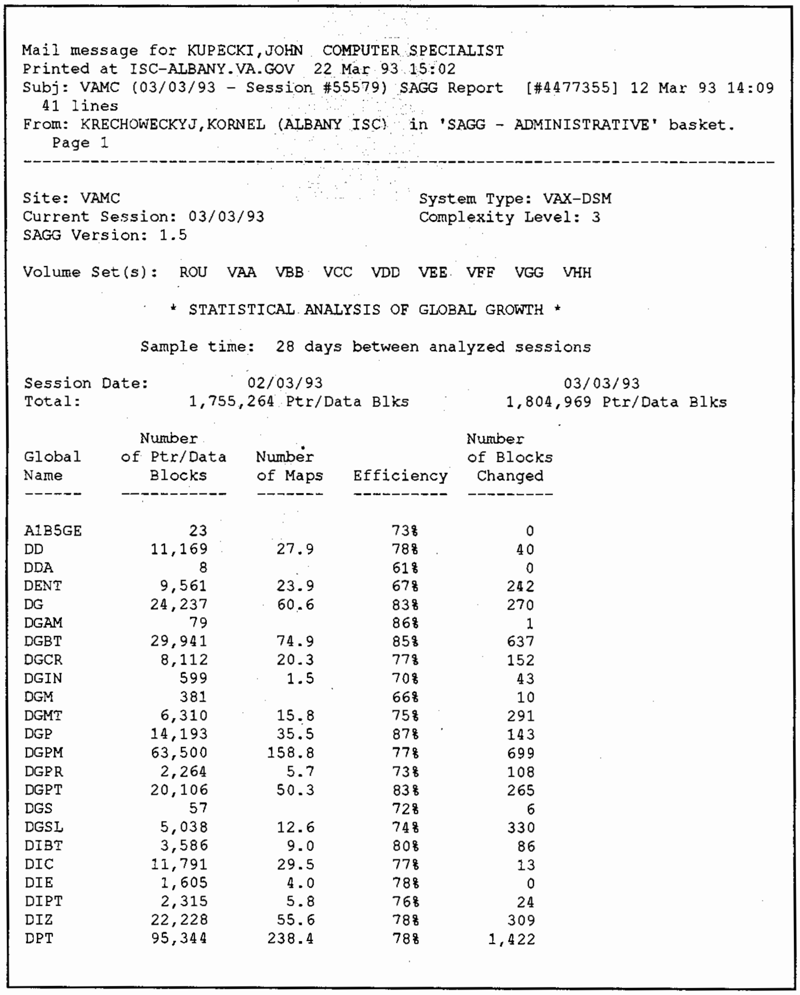
Once the SAGG collection cycle has completed, a local electronic mail message is produced and delivered to the local SAGG mail group. If the cycle did not properly complete, an electronic mail message is generated to warn the computer center operation's staff of a problem.

After the data collection cycle has completed at the medical center, the temporary data collection global is immediately moved into a network mail message. This information is then automatically transferred to the centralized SAGG database over network mail by the VA developed MailMan utility. Once the network message arrives, MailMan delivers the incoming message to an automated 'server' utility. This server is designed to automatically manipulate message data from MailMan without user intervention. The server software accomplishes several important functions. First, the server generates an electronic global summary mail message which is sent back to the local SAGG mail group at the participating site through network mail. This message gives the site immediate feedback about global growth statistics during the captured session. Next, the server merges the incoming global information from the site into the centralized SAGG database where further statistical analyses are performed. The server accomplishes this merge by parsing through the mail message which contains the structure and contents of the temporary collection global from the site. Each node of this global is inserted into the appropriate field of the centralized SAGG database.

SAGG STATISTICAL REPORTS

After each data collection cycle, an electronic global summary mail message is immediately transmitted back to the originating site through network mail. This message contains a summary of many pertinent aspects including global block size, global percent efficiency, change in global size between sampling sessions with explanatory footnotes, globals resident on each production volume set, total database size and growth difference between sampling sessions, and miscellaneous operating system information. Figure 1 represents a partial example of a typical global summary message from a participating site.

After the global information from a site has been merged into the centralized SAGG database, further statistical analyses are conducted. Examination employs both MUMPS and VA developed FileMan utilities. Also, some data is moved into Microsoft EXCEL spreadsheets for further analysis. All analyses are summarized on a monthly basis and are distributed to the DHCP administrators for their review. Part of this analysis entails the intentional grouping of the medical centers into four subsets called Complexity Levels. These subsets are based on factors which are independent of the medical center's database size and growth and is done to preclude the generation of averaging abnormalities between multi-divisional hospitals. The monthly report contains many individual tables and graphs summarizing the cumulative global information that was captured and analyzed during the particular month.



```
Mail message for KUPECKI,JOHN  COMPUTER SPECIALIST
Printed at ISC-ALBANY.VA.GOV  22 Mar 93 15:02
Subj: VAMC (03/03/93 - Session #55579) SAGG Report  [#4477355] 12 Mar 93 14:09
  41 lines
From: KRECHOWECKYJ,KORNEL (ALBANY ISC)  in 'SAGG - ADMINISTRATIVE' basket.
   Page 1
-----------------------------------------------------------------------------

Site: VAMC                                System Type: VAX-DSM
Current Session: 03/03/93                 Complexity Level: 3
SAGG Version: 1.5

Volume Set(s):  ROU  VAA  VBB  VCC  VDD  VEE  VFF  VGG  VHH

               * STATISTICAL ANALYSIS OF GLOBAL GROWTH *

            Sample time:  28 days between analyzed sessions

Session Date:          02/03/93                         03/03/93
Total:           1,755,264 Ptr/Data Blks          1,804,969 Ptr/Data Blks

             Number                               Number
Global     of Ptr/Data   Number                  of Blocks
Name         Blocks      of Maps    Efficiency    Changed
------     -----------   -------    ----------   ---------


A1B5GE          23                     73%           0
DD          11,169         27.9        78%          40
DDA              8                     61%           0
DENT         9,561         23.9        67%         242
DG          24,237         60.6        83%         270
DGAM            79                     86%           1
DGBT        29,941         74.9        85%         637
DGCR         8,112         20.3        77%         152
DGIN           599          1.5        70%          43
DGM            381                     66%          10
DGMT         6,310         15.8        75%         291
DGP         14,193         35.5        87%         143
DGPM        63,500        158.8        77%         699
DGPR         2,264          5.7        73%         108
DGPT        20,106         50.3        83%         265
DGS             57                     72%           6
DGSL         5,038         12.6        74%         330
DIBT         3,586          9.0        80%          86
DIC         11,791         29.5        77%          13
DIE          1,605          4.0        78%           0
DIPT         2,315          5.8        76%          24
DIZ         22,228         55.6        78%         309
DPT         95,344        238.4        78%       1,422
```

Figure 1. Example of Partial Global Summary Message

Continued on page 132

Table 1 shows the average database size, growth, and percent growth summary data for each of the four Complexity Levels for each month starting from February 1992 to the current sample month. The global size information from this table is depicted in graphical form in Figure 2.

Another section of the monthly report shows tabular and graphical representations of the largest DHCP software packages for the four Complexity Levels. Each package is comprised of a certain number of related globals. Figure 3 graphically depicts both package size and percentage information for Complexity Level 1 sites during a sample month.

Also shown in one of the spreadsheets is a variety of package and global information. Global data for each site is listed by Complexity Level and is arranged by individual DHCP package. The report uses various printing fonts and shading to furnish additional information such as successful data transmission, dynamic growth capabilities, purging and archiving abilities, and standard deviation variances above or below the mean. Table 2 shows an example of this spreadsheet for one group of Complexity Level sites.

BENEFITS OF THE SAGG PROJECT

Use of the SAGG Project produces beneficial information for the medical centers and other organizations within the VA. Some noted advantages are improved database management and a greater ability to trend database growth within the evolving DHCP program. At the medical centers, SAGG data has been integral in many system upgrade and tuning studies. Sites are supplied with detailed listings on a regular basis showing database size and growth rates. Also shown are any duplicate globals, global efficiency percentages, guidance on which globals are above or below the statistical mean for their Complexity Level, and informational footnotes stating possible reasons. When supplied with this information, the site can make informed decisions on evaluating current purging and archiving schedules, forecasting for equipment needs, balancing data access across systems to effect better resource utilization, and allocating resources.

Table 1. Total Database Growth Summary by Complexity Level

COMPL LEVEL	FEB AVG SIZE (megabytes)	Part A. Size Summary (megabytes)									
		MARCH AVG SIZE	APRIL AVG SIZE	MAY AVG SIZE	JUNE AVG SIZE	JULY AVG SIZE	AUG AVG SIZE	SEPT AVG SIZE	OCT AVG SIZE	NOV AVG SIZE	DEC AVG SIZE
1	3,385.19	3,704.84	3,819.54	3,930.89	4,120.55	4,291.30	4,499.10	4,551.58	4,813.68	4,754.49	4,785.92
2	2,568.68	2,462.24	2,698.30	2,709.91	2,804.02	2,840.03	2,935.97	3,114.10	3,173.68	3,250.92	3,219.01
3	1,567.43	1,471.22	1,594.66	1,657.54	1,629.61	1,642.09	1,689.81	1,886.65	1,925.68	1,979.22	1,952.97
4	1,177.44	1,148.26	943.96	930.36	938.02	916.29	1,062.29	1,212.21	1,247.12	1,331.25	1,346.43

COMPL LEVEL	FEB AVG SIZE (megabytes)	Part B. Growth Summary (megabytes)											
		FEB-MAR GROWTH	MAR-APR GROWTH	APR-MAY GROWTH	MAY-JUNE GROWTH	JUN-JUL GROWTH	JUL-AUG GROWTH	AUG-SEPT GROWTH	SEPT-OCT GROWTH	OCT-NOV GROWTH	NOV-DEC GROWTH	TOTAL GROWTH FEB-DEC	AVG MONTHLY GROWTH FEB-DEC
1	3,385.19	319.65	114.70	111.35	189.66	170.75	207.80	52.48	262.10	-59.19	31.43	1,400.73	140.07
2	2,568.68	-106.44	236.06	11.61	94.11	36.01	95.95	178.13	59.58	77.24	-31.91	650.33	65.03
3	1,567.43	-96.21	123.44	62.88	-27.93	12.48	47.72	196.84	39.03	53.54	-26.25	385.54	38.55
4	1,177.44	-29.18	-204.30	-13.60	7.66	-21.73	146.00	149.92	34.91	84.13	15.18	168.99	16.90

COMPL LEVEL	FEB AVG SIZE (megabytes)	Part C. Percent Growth Summary											
		FEB-MAR % GROWTH	MAR-APR % GROWTH	APR-MAY % GROWTH	MAY-JUNE % GROWTH	JUN-JUL % GROWTH	JUL-AUG % GROWTH	AUG-SEPT % GROWTH	SEPT-OCT % GROWTH	OCT-NOV % GROWTH	NOV-DEC % GROWTH	TOTAL % GROWTH FEB-DEC	AVG MONTHLY % GROWTH FEB-DEC
1	3,385.19	9.44%	3.10%	2.92%	4.82%	4.14%	4.84%	1.17%	5.76%	-1.23%	0.66%	41.38%	3.56%
2	2,568.68	-4.14%	9.59%	0.43%	3.47%	1.28%	3.38%	6.07%	1.91%	2.43%	-0.98%	25.32%	2.34%
3	1,567.43	-6.14%	8.39%	3.94%	-1.69%	0.77%	2.91%	11.65%	2.07%	2.78%	-1.33%	24.60%	2.34%
4	1,177.44	-2.48%	-17.79%	-1.44%	0.82%	-2.32%	15.93%	14.11%	2.88%	6.75%	1.14%	14.35%	1.76%

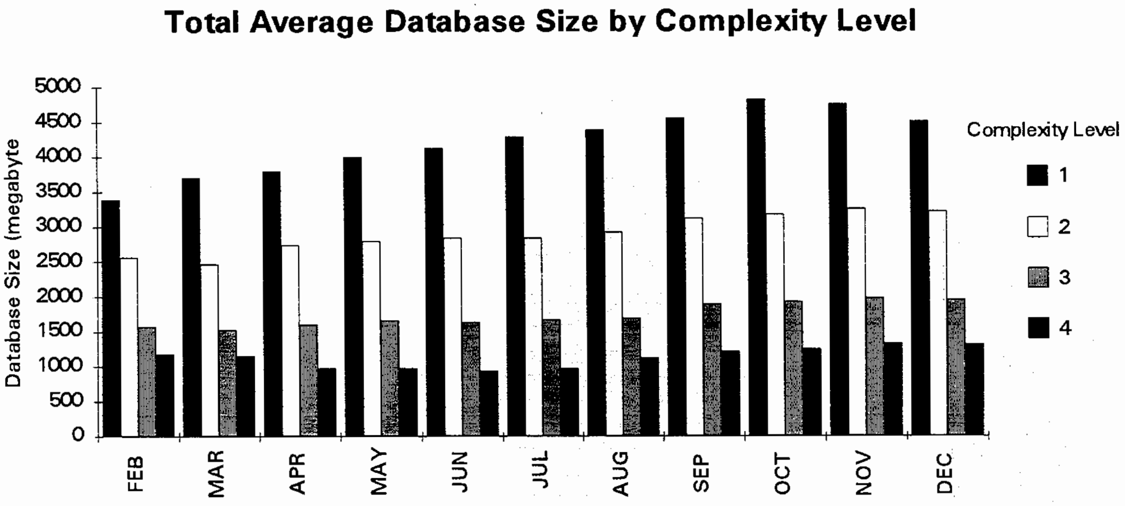


Figure 2. Total Average Database Size by Complexity Level

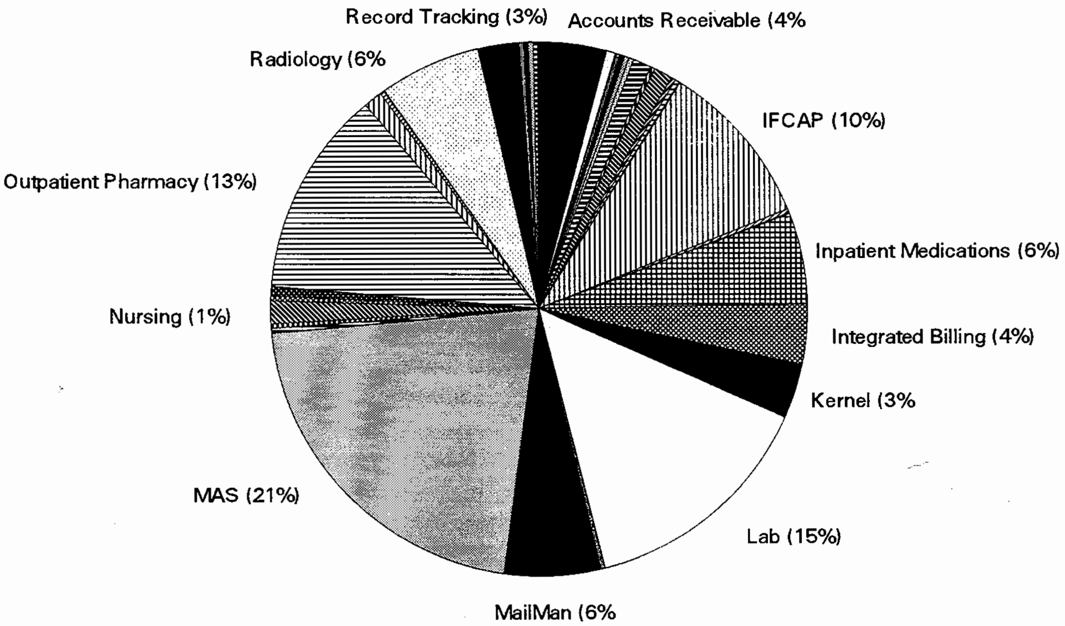


Figure 3. DHCP Package Size Percentages for Complexity Level 1

Table 2. Global Sizes for Sites by Complexity Level

		CORE + 4 PACKAGES					(data in blocks)	
		ACCOUNTS. RECEIVABLE	DIETETICS					
	COMPL	PRCA	FH	FHEN	FHING	FHNU	FHPT	FHUM
ISC 1:								
SITE 1	1	281,055	564	2,205	145	1,582	19,064	150
SITE 2	1	147,707	373	1,757	129	1,241	23,432	106
SITE 3	2	159,763	791	5,694	150	1,356	39,344	321
SITE 4	2	151,026	382	3,700	142	1,552	14,177	312
SITE 5	2	220,164	264	425	145	1,633	30,354	1,008
SITE 6	2	161,255	1,256	4,210	164	1,635	33,449	90
SITE 7	3	76,980	304	1,122	129	1,270	13,062	98
SITE 8	3	91,886	458	2,696	121	1,214	2,424	66
SITE 9	4	90,724	345	292	138	1,229	4,006	192
SITE 10	4	101,00[illegible]	5,586	2,396	137	1,235	14,867	222
SITE 11	4	41,819	467	2,077	148	1,232	12,972	86

Dark shading: 1 std. dev. ABOVE mean (by comp. level)
Light shading: 1 std. dev. BELOW mean (by comp. level)
Bold type: Dynamic - Italics: Purgeable and/or Archivable

DHCP administrators are using SAGG reports to enhance site support activities. Trending analysis helps to recognize and predict any emerging problems that may lead to operational difficulties. Evidence is also gained which may lead to improved system tuning and management methodologies. Similarly, reliable statistics provide information concerning the size and expected growth of various DHCP packages and the merits of present purging and archiving strategies. A recently completed study showed significant discrepancies between disk requirement estimates based on current package sizing model algorithms and actual disk space usage based on SAGG data. Careful study is continuing to ascertain the causes of these discrepancies and to supply updated algorithms which may prove more accurate.

The SAGG Project facilitates an objective analysis of current database information to assist in procurement strategies and managing the DHCP program as a whole. SAGG data is also being used to validate the prediction of the DHCP model for disk space needed to support the DHCP program. While still in progress, these results may help to improve upon present sizing methodologies so that resources can be efficiently allocated.

CONCLUSION

The SAGG Project has become a vital statistical tool in the DHCP program. Maintaining a current and reliable database on global growth characteristics offers the ability to recognize and correct emerging problems before they cause operational difficulties. Greater knowledge can lead to the development of additional management tools. Management of DHCP is improved from the site level through the program administrator level by more efficient use of computer resources. Planning and strategies are enhanced through greater informed decision making, resulting in cost savings. Trending analysis identifies the impact of package implementation and the actual effects of purging and archiving. Future trending functionality will show the relationship between package, global, and file activity. Similarly, ongoing analysis will allow continual evaluation of DHCP sizing model accuracy and new ways of interpretation. Adapting this approach may lead to a different manner of medical center grouping to better define their computer resource needs.