Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

USE OF CLASS STORAGE ESTIMATION TOOL FOR CAPACITY PLANNING

Michael E. Fotta

Global Science & Technology, Inc.
2000 Green River Drive
Fairmont WV 26554, USA

ABSTRACT

The Comprehensive Large Array-data Stewardship System (CLASS) archives environmental data from many NOAA sources and NOAA users. CLASS typically charges individual NOAA customers for this storage. Users of the CLASS system desired a capacity estimation tool that would enable them to easily make estimates of the cost before committing to use CLASS to archive their data. Furthermore, users wanted to be able to manipulate the values of variables related to this storage which were under their control; that is, variables that could increase or lower the cost. Forio's Simulate™ was used to develop a web-based capacity planning simulation - the CLASS Storage Estimation Tool (CSET) - to meet these needs.

1 INTRODUCTION

The Comprehensive Large Array-data Stewardship System (CLASS) archives over 6.5 petabytes of NOAA's environmental data and is adding to these holdings at a current rate of over 1.5 PB per year. With the addition of new major data sources this growth will increase to over 7 PB per year by the end of 2016. In order to process, archive and deliver this data CLASS uses Storage Area Network (SAN) disks in a configuration of multiple temporary storage caches and a tape library. Typically 1 TB/day of data ingested needs about 44 TB of disk cache for processing, 15 TB of cache for delivery and 146 tapes per year.

It is critical that CLASS has the ability to estimate the storage media necessary for future data. Overestimating leads to over spending, but underestimating leads to a shortfall in the storage available for new data. Estimates are needed for projecting both the overall growth of CLASS storage needs and the storage needs for individual campaigns (particular data sets). The latter is especially true as CLASS charges the individual NOAA campaigns for the media resources used. It is not surprising then that users of the CLASS system desired a capacity estimation tool that would enable them to make estimates of the cost for their particular campaign. Furthermore, users wanted to be able to run multiple simulations enabling them to manipulate variables related to this storage which were under their control; that is, variables that could increase or lower the cost. In order to meet these needs Forio's Simulate™ was used to build the CLASS Storage Estimation Tool (CSET).

2 INPUT VARIABLES AND USE OF CSET

The data input variables which determine the specific media needs and are under a user's control are: 1) daily data ingest rate to CLASS, 2) the percent of data to be kept permanently on SAN disks, 3) the volume of their data to be delivered from CLASS storage daily, 3) the number of files ingested daily, 4) the size of these files, 5) the period of time (months or years) to ingest the data, and 6) the "situation" for ingesting the data into CLASS. The situation refers to how the data is sent to CLASS - from either Backlog (an existing store of data), Operational (on a daily basis), or a combination of the two. Also, the Operational data variables may change on an annual basis. This leads to five different situations for ingesting data: 1) Backlog only, 2) Operational – no input variation, 3) Operational with yearly input variation, 4) Backlog and Operational with no input variation, and 5) Backlog and Operational with yearly variation.

The CSET model enables users to run simulations for these different situations while manipulating values for these variables in order to determine the media capacities and cost for each simulation run, and then compare these runs. A campaign often has some discretion in how the data is brought in, Backlog and/or Operational, and in the variables characterizing the input of the data as described above. Users often start with variable values that would give them the fastest ingest of the largest number of files per day. After a user enters their values and runs the simulation CSET provides a summary Resource Estimation web page for that situation. If the costs are greatly outside their budget they may seek another avenue for archiving their data instead of CLASS.

Alternatively they can go back and run a new simulation with reduced values for some variables (e.g., reduce the daily ingest rate). They then save the values for the input variable used and results from the Resource Estimation for this run for comparison with other simulation runs. The data from the saved runs is then presented in a Run Table as discussed in the next section. Although some use of CSET has been made for long term planning, the more common use is by users determining how changing one or more variables affects capacities and cost for their campaign.

3 COMPARING SIMULATION RUNS

Once a user has executed a number of runs with different variable inputs they can use the Run Tables to get an understanding of how their changes affect media capacity and cost. As the name implies a Run Table presents the data for each saved run. In order that users get the maximum information possible, but presented in a fashion that does not overwhelm them, multiple Run Tables are viewed in a web page with separate tabs (see Figure 1). There are tabs for Final Results, Backlog Results, Operational Results and Inputs Run Tables. The Final Results are equivalent to the Backlog or Operational Results when data is ingested in only one of these ways. However, the Final Results are different, and not a simple combination, when campaign data is ingested in both fashions. Presenting the Final Results separately also enables a user to determine whether the Backlog or Operational data is the capacity and cost driver.

FINAL | BACKLOG | OPERATIONAL | INPUTS

RunName	Ingest Rate	SFS Capa...	SFS Disks	HPSS Cap...	HPSS Disks	Tape Capa...	Tapes	Drives	Disk Cost	Tape Cost	Drive Cost	Total Cost
Run 4	1.1	60.17	21	13.53	4	1,642.5	657	1	$96,000	$65,700	$20,000	$181,700
Run 3	1.3	193.02	65	15.99	4	1,642.5	657	1	$272,000	$65,700	$20,000	$357,700
Run 2	1.1	132.44	45	13.53	4	1,642.5	657	1	$192,000	$65,700	$20,000	$277,700
Run 1	1.5	180.6	61	18.45	5	1,642.5	657	1	$259,000	$65,700	$20,000	$344,700

Figure 1. Example CSET Run Table Web Page.

Using the Run Table users can compare the disk, tape and drive capacities and cost resulting from their variable inputs on different Runs. By visually comparing the results and using the Input tab users can determine how changes in a variable or variables affect these capacities and costs. All of the data shown in the Run Tables can also be saved to Excel or other tools for further analysis, if so desired.

In the example shown in Figure 1 the user did four runs and noticed that variation in Disk Cost is the main driver in changes in Total Cost. Looking through the data it can be seen that Disk Cost variation appears related to the variation in SFS Disks as the HPSS Disks hardly vary. The user then saw that the Ingest Rate was the same (1.1) for the two lowest number of SFS Disks, but the Run 4 had a much lower number of SFS Disks than Run 2. At this point the user would refer to the Inputs tab to discover how the inputs varied. In this case the user found that Run 4 had only specified that 1% of data be kept permanently on disks while the other run had specified 10%. The user decided that the reduced use of disks was acceptable given the reduction in cost.

Detailed examples of how users have applied the CSET, including how variables are entered and modified, and Runs compared will be presented.