Paper 133-2008

Data Integration in a Grid-Enabled Environment

Cheryl Doninger, SAS Institute Inc., Cary, NC
Gary Mehler, SAS Institute Inc., Cary, NC
Nancy Rausch, SAS Institute Inc., Cary, NC

ABSTRACT

SAS® Data Integration Studio and SAS® Grid Manager add capabilities to the SAS® product suite to distribute workloads across a grid of computers and thereby allow large processes to complete more quickly than previously possible. SAS Grid Manager has been incorporated into SAS Data Integration Studio to facilitate using grid resources for any long-running task that can be processed in parallel to another task. This paper discusses typical data integration workloads, how to scale them on typical grid computing hardware, and the new capability to load balance multiple data integration tasks across grid resources.

INTRODUCTION

SAS was originally developed as a 4GL programming language that could be used to write SAS applications. These applications executed in a single, sequential path that matched the single CPU capabilities at that time.

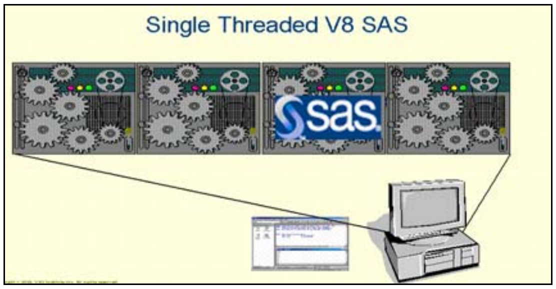


Figure 1. SAS Version 8 Processing Capabilities

As data volumes continued to grow and computing needs continued to increase, hardware vendors responded by creating symmetric multi-processor (SMP) machines. SMP is a multiprocessor computer architecture in which two or more identical processors are connected to a single, shared main memory. Organizations also built networked computing environments with multiple, individual computing resources connected via networking protocols. SAS responded to these advances by developing multi-processing capabilities. In SAS/CONNECT Version 8, SAS applications could take advantage of the multi-processors available in desktop and server platforms, and could multi-process across platforms available in a network. With SAS/CONNECT you can spawn N SAS sessions or processes, simultaneously execute N tasks as independent processes, and coordinate the execution and results into the client or parent session. A major benefit of this technology is the flexibility for the multiple sessions to run on multiple CPUs within an SMP box, across multiple, distributed machines in a network, or a combination of both.

SAS Version 9 brought multi-threading capabilities to further leverage the growing adoption of SMP architectures. On hardware with more than one CPU, multi-threading provides a mechanism for a program to exploit more than one CPU simultaneously. By creating multiple, simultaneously active threads, the program enables the operating system to schedule these threads concurrently on more than one CPU. Several SAS procedures, including SORT, SUMMARY, REG, and GLM, as well as several SAS servers and SAS/ACCESS engines were rewritten with multi-threading capabilities. It is important to understand that while multi-threading capabilities allow an application to exploit the multiple CPUs in an SMP machine, the scalability gains are limited to a single SMP box and cannot leverage distributed computing resources.

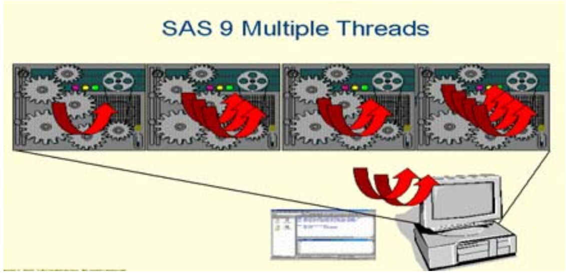


Figure 2. SAS V9 Processing Capabilities

The initial offering of SAS Version 9 provided both multi-threading and multi-processing capabilities to allow SAS applications to scale-up to take advantage of multi-processors available in SMP hardware. SAS also provided multi-processing capabilities to allow SAS applications to scale-out to take advantage of any number of distributed computing resources. The next evolutionary step was for SAS to make it possible for customers to run their SAS applications in a grid or cluster environment. SAS Grid Manager was introduced in SAS 9.1.3 to build upon the parallel capabilities of SAS/CONNECT and to add the many other requirements of enterprise grid deployments. SAS Grid Manager provides multi-user load balancing, policy enforcement, efficient resource allocation, and prioritization for SAS products and solutions running in a shared grid environment.

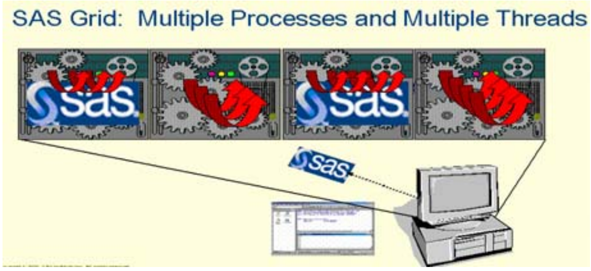


Figure 3. SAS V9.1.3 Can Leverage Grid Processing

SAS Grid Manager has been integrated with many SAS products and solutions to provide seamless grid capabilities to the users of these applications. One such application is SAS Data Integration Studio. This paper will discuss the multiple ways that SAS Data Integration Studio works with SAS Grid manager to bring the benefits of a SAS grid infrastructure to the user in an easy-to-use, point-and-click development environment.

USING SAS GRID MANAGER IN SAS DATA INTEGRATION STUDIO

When running processes in a grid, SAS Grid Manager dynamically determines node availability and monitors grid nodes to determine which node is the best candidate to receive the next workload segment. This determination can be based on many factors, but it often considers the current load under which all grid nodes are running at any given time. The node that has the lowest CPU load becomes the best candidate on which to run the next workload segment. This dynamic capability greatly increases job runtime performance by distributing processes across a wider array of resources capable of handling the greater computing load.

There are three key ways that SAS programs and applications can exploit grid computing using SAS Grid Manager in SAS Data Integration Studio:

- Distributed Enterprise Scheduling—distribute scheduled jobs to a shared pool of resources in a grid
- Multi-user Workload Balancing—distribute user-submitted jobs interactively to a shared pool of resources in a grid
- Parallel Workload Balancing—distribute parallelized jobs, either scheduled or user submitted, to a shared pool of resources in a grid

Figure 4 below describes the SAS products that can exploit each of the features. SAS Data Integration Server can exploit all three capabilities.

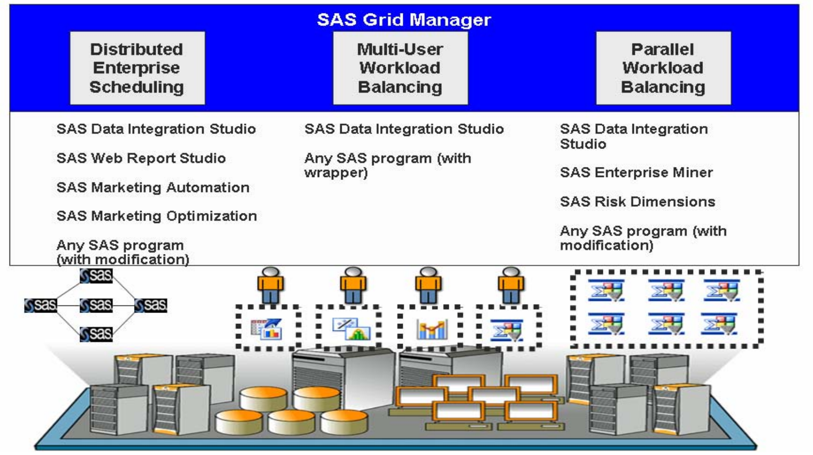
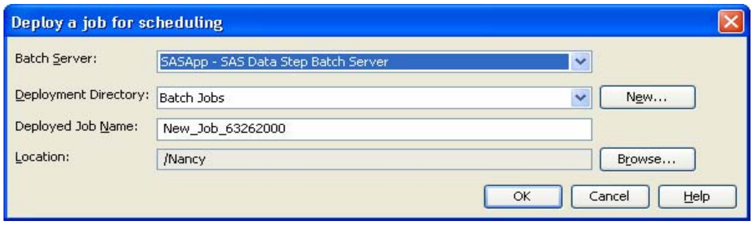

SAS Grid Manager
Distributed Enterprise Scheduling
Multi-User Workload Balancing
Parallel Workload Balancing
SAS Data Integration Studio
SAS Web Report Studio
SAS Marketing Automation
SAS Marketing Optimization
Any SAS program (with modification)
SAS Data Integration Studio
Any SAS program (with wrapper)
SAS Data Integration Studio
SAS Enterprise Miner
SAS Risk Dimensions
Any SAS program (with modification)
sas

Figure 4. Products in SAS That Can Exploit SAS Grid Computing

Distributed Enterprise Scheduling

Using Distributed Enterprise Scheduling with SAS Grid Manager, scheduled jobs are targeted to run on the grid. This makes all resources in the grid available to run scheduled jobs. The scheduling server manages the resources in the grid workload so that jobs are efficiently distributed across the available machines in the grid. You can leverage this capability in SAS Data Integration Studio and other SAS products by deploying the job to be scheduled and then using the scheduling server that manages the grid to schedule and run the jobs.



Deploy a job for scheduling

Batch Server: SASApp - SAS Data Step Batch Server

Deployment Directory: Batch Jobs

Deployed Job Name: New_Job_63262000

Location: /Nancy

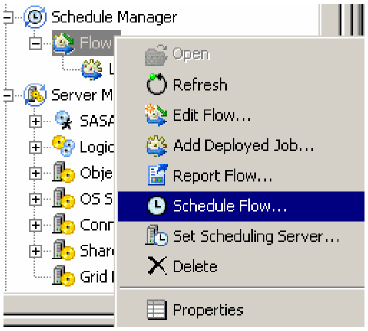


Figure 5. Using Schedule Manager for Grid Computing

Multi-User Workload Balancing

With Multi-User Workload balancing, individual users have the ability to submit jobs interactively and directly to the grid. For example, a site could have a number of users that do ad hoc development, such as model development, queries, and other sorts of discovery and analysis. SAS Grid Manager provides the ability to leverage the grid when submitting jobs. Using the grid provides all of the capabilities of load balancing, such as queuing, prioritization, workload balancing, and resource management, for this type of interactive submit. High priority jobs can even preempt lower priority work so that the most critical business processes execute first. This enables users to leverage all of the available resources in their distributed environment for job processing, thereby speeding up long-running tasks and increasing user productivity.

SAS Data Integration Studio 4.2 supports this capability by enabling you to select the target server where you want to submit jobs or transformation steps in a job. The target server can be the grid server. When selected, Data Integration Studio wrappers the submitted code with the appropriate statements to submit the interactive job to the grid.

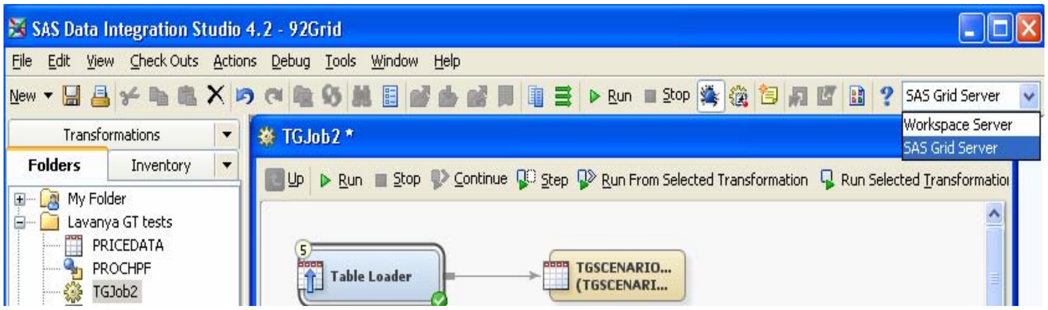
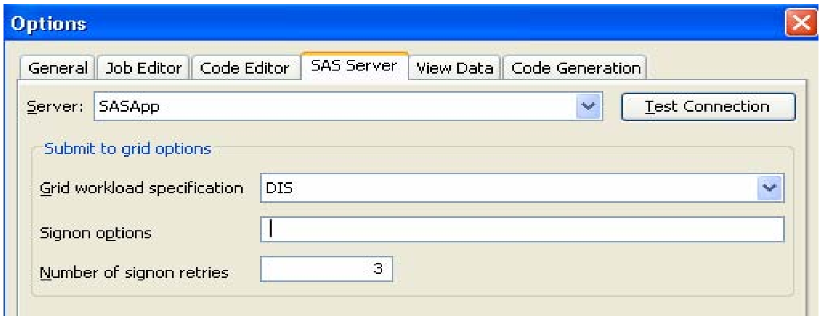


Figure 6. Using SAS Data Integration Studio's server selection capabilities to submit jobs to a Grid

SAS Grid Manager also supports the option to group like resources together into a group of nodes. For example, an administrator might want to configure one set of nodes for work with analytical applications, and another set of nodes for data integration processes. The administrator can assign a name to this partition so that users can specify it when submitting their processes. This partitioning ability allows the administrator to tailor the grid to better meet the needs of the user community. Data Integration Studio allows you to specify an optional grid partition when you submit processes using Workload balancing.



Server: SASApp

Submit to grid options

Grid workload specification: DIS

Signon options:

Number of signon retries: 3

Figure 7. Data Integration Studio Supports Grid Partitions

Parallel Workload Balancing

SAS Data Integration Studio also supports the ability to parallelize processes, which can be submitted interactively using Parallel Workload balancing or can be scheduled as job flows using Distributed Enterprise Scheduling to run in parallel on a grid. Parallel execution of job flows is supported using iteration, a scenario frequently found in Data Integration processing. Iteration can be explained with a simple example.

Sometimes it is desirable to execute the same process flow over and over again on different data. For example, suppose you have United States Census data as a set of 50 tables, one for each state. The table structure is identical, but the data is specific to each state:

- HouseholdsCA
- HouseholdsTX
- HouseholdsAZ
- HouseholdsNM

…additional tables

Now suppose you want to calculate the number of households that own more than one acre of land. For each state of data you would run the same process on every source table. One way to process this data would be to run the process one run at a time per table. However, if you have the computing resources that would be available using a grid, you could run the same process in parallel on the different source tables. This is iteration, which means that the same process flow is run iteratively for each input in sequence. The scope of the iteration is called a loop. SAS Data Integration Studio includes a methodology for handling looping for jobs, both serially and in parallel. Each run is called an iteration of the job, and it can be submitted to run in parallel, either on a single machine, or submitted to a grid to leverage the multiple computing nodes available on a grid.

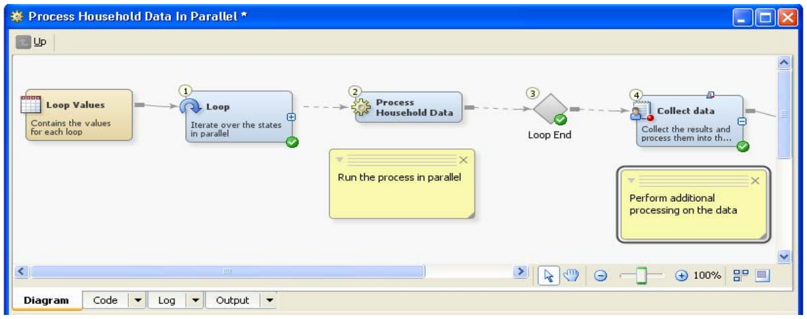
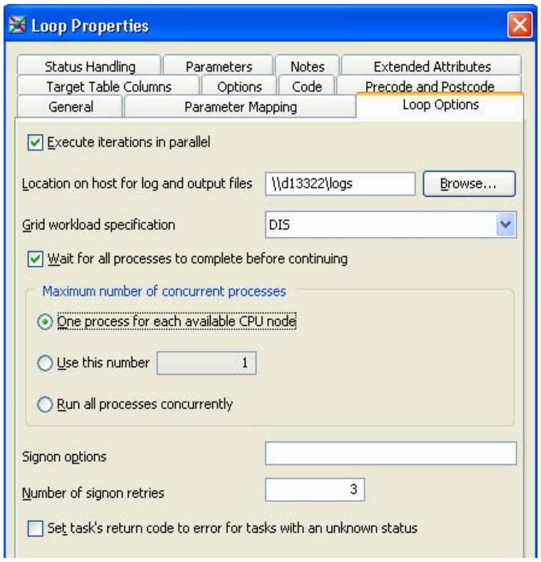

Process Household Data In Parallel *
Up
Loop Values
Contains the values for each loop
1
Loop
Iterate over the states in parallel
2
Process Household Data
3
Loop End
4
Collect data
Collect the results and process them into th...
Run the process in parallel
Perform additional processing on the data
100%
Diagram
Code
Log
Output

Figure 8. SAS Data Integration Studio Job That Supports Parallel Processing of Data, One Process per Loop

The loop transform lets you specify submit options to optimize the code to match the hardware that you will be submitting to. These options include dynamic discovery of available CPUs, optional grid workload configuration so that job submissions are sent only to the right set of available servers, sign-on retries, and other features.



☑ Execute iterations in parallel

Location on host for log and output files: \\d13322\logs

Grid workload specification: DIS

☑ Wait for all processes to complete before continuing

Maximum number of concurrent processes

● One process for each available CPU node

○ Use this number: 1

○ Run all processes concurrently

Signon options:

Number of signon retries: 3

□ Set task's return code to error for tasks with an unknown status

Figure 9. Options Available in the Loop Transform for Iteration

PERFORMANCE CONSIDERATIONS

It is useful to understand how a grid and parallelization can improve job time performance. Taking the census example, some states have a larger number of households than other states. The performance of each job element therefore varies by state. We took a baseline test and benchmarked the runtimes for each state during sequential execution to come up with the following performance workload running in a test environment of 10 blade servers with 60GB of data.

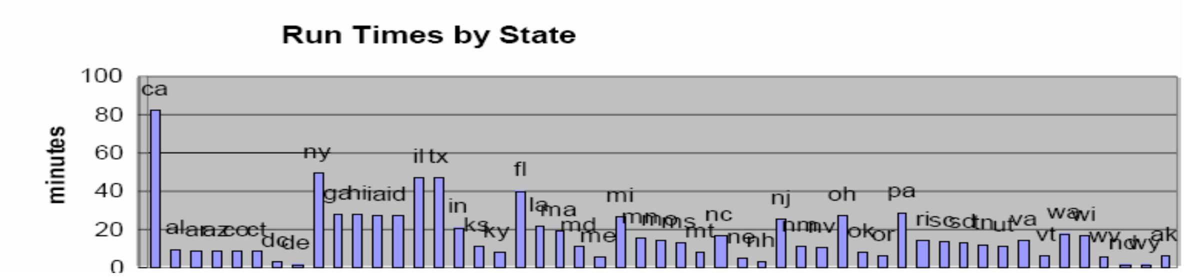
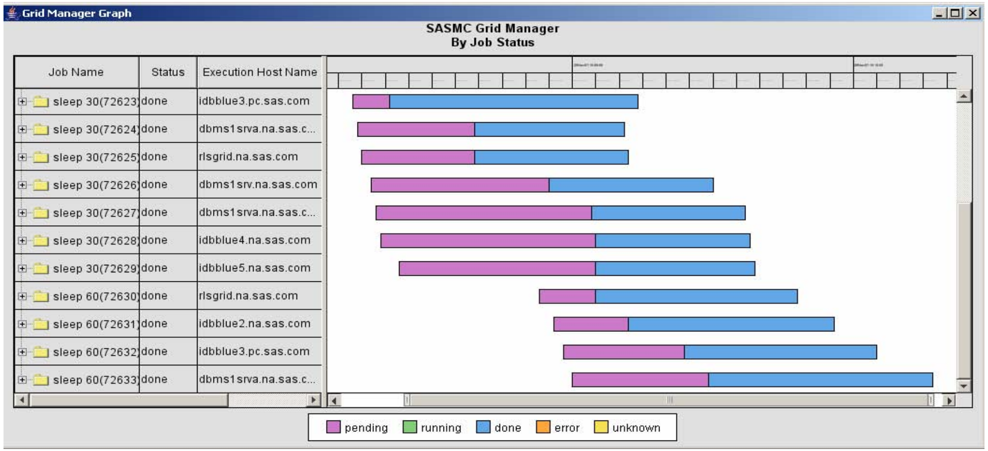


Figure 10. Runtimes for Processing Census Data by State

Running these jobs serially took around 600 minutes of real time on our test environment. We then ran the same jobs in parallel using the Loop transform. We were able to achieve a best-case runtime in the test environment of approximately 100 minutes, a 6x performance gain over the serial case.

SAS has published a number of grid performance benchmarks for different scenarios, see the reference list at the end of this paper for details. Benchmark tests are available for a variety of customer-usage patterns, including a large, multi-user, ad hoc analytics environment and an I/O intensive scenario. These scenarios varied the number of available grid nodes and I/O capabilities to determine performance patterns with some encouraging results. In the computational scenario, the benchmarks were able to achieve linear scalability as work load increased simply by increasing the available computer resources, that is, adding nodes into the grid. Similarly with I/O intensive processes, increasing the available I/O resources enabled the processes to scale linearly in the grid environment.

When using the grid for testing, we found it useful to visualize the workload on the grid to identify when processes were running and when they had completed. This is possible using the Grid monitoring capabilities of the Grid Manager plug-in available in SAS Management Console.



Job Name	Status	Execution Host Name
sleep 30(72623)	done	idbblue3.pc.sas.com
sleep 30(72624)	done	dbms1srva.na.sas.c...
sleep 30(72625)	done	rlsgrid.na.sas.com
sleep 30(72626)	done	dbms1srv.na.sas.com
sleep 30(72627)	done	dbms1srva.na.sas.c...
sleep 30(72628)	done	idbblue4.na.sas.com
sleep 30(72629)	done	idbblue5.na.sas.com
sleep 60(72630)	done	rlsgrid.na.sas.com
sleep 60(72631)	done	idbblue2.na.sas.com
sleep 60(72632)	done	idbblue3.pc.sas.com
sleep 60(72633)	done	dbms1srva.na.sas.c...

Figure 11. Grid Monitoring Capabilities Available in SAS Grid Manager

Other Grid Manager features enable you to see the status of each job process on the grid, identify which nodes are in use and which nodes are available, withdraw a grid node from the set of available nodes, and perform other useful operations. You can tune your grid environment to achieve maximum efficiency.

TEXT MINING SCENARIO

Parsing unstructured data is one way you can readily leverage grid computing in the data integration environment. In some organizations, unstructured data, mostly in the form of text files, can account for a significant amount of useful information in an organization's knowledge stores. This information is difficult to access because it is not easily synthesized into tables that can be merged into existing processes, and the volume of information can require a significant amount of time to process. Grid processing, with its ability to manage workloads and bring many resources to bear, is ideally suited to handling this type of data.

SAS Text Miner includes capabilities to assist in processing unstructured text data. You can wrapper the text mining calls into Data Integration Studio transformations to process unstructured data coming from text files. For example, suppose you want to process text data coming from a web page, and merge that data with other structured data in your data warehouse. Follow this process:

1. Generate a wordlist from a website.
2. Process the word list.
3. Merge the structured and unstructured data.

In Data Integration Studio we can create transforms that call SAS Text Miner to process the data on the web pages.

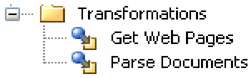
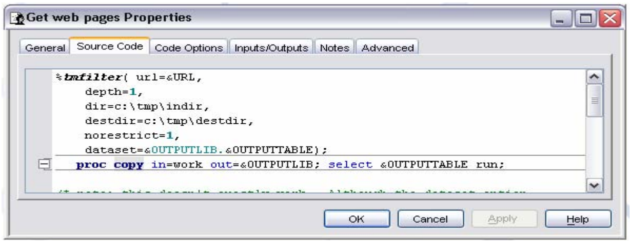


Figure 12. Transforms in Data Integration Studio for Transforming Unstructured Text Data

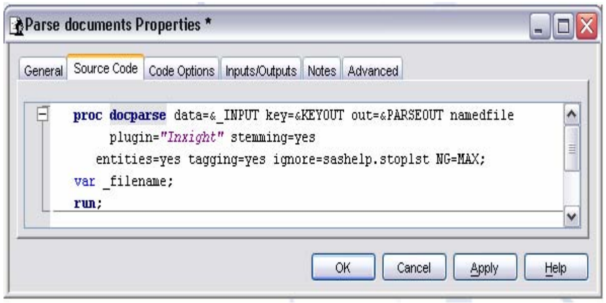
The Get web pages transform uses the text miner %TMFILTER macro.



```
%tmfilter( url=&URL,
    depth=1,
    dir=c:\tmp\indir,
    destdir=c:\tmp\destdir,
    norestrict=1,
    dataset=&OUTPUTLIB.&OUTPUTTABLE);
  proc copy in=work out=&OUTPUTLIB; select &OUTPUTTABLE run;
```

Figure 13. Source Code Example from the Get Web Pages Transform in SAS Data Integration Studio

This transform reads web pages starting from a user-specified URL and creates a data set that can be used as input. Documents of different formats can also be read using this same technique. Then we parse the documents, which creates word lists by type and frequency. Parts of speech can also be identified during the parsing stage.



```
proc docparse data=&_INPUT key=&KEYOUT out=&PARSEOUT namedfile
     plugin="Inxight" stemming=yes
   entities=yes tagging=yes ignore=sashelp.stoplst NG=MAX;
var _filename;
run;
```

Figure 14. Source Code Example from the Parse Documents Transform in SAS Data Integration Studio

Using these two transforms, we can read web pages and parse them into two data sets, a list of keywords found and the raw-word data.

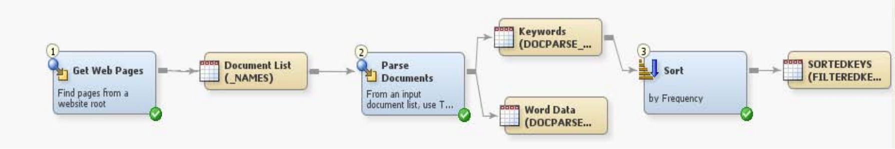


Figure 15. Sample Data Integration Studio Job To Parse Text Data

Using these two data sets, we can further process the data to pull out standard roles from the raw list of words, apply a standardization rule to place the words into known sets, and match the words to the roles so that we have a set of data that is structured to match up to additional warehouse data.

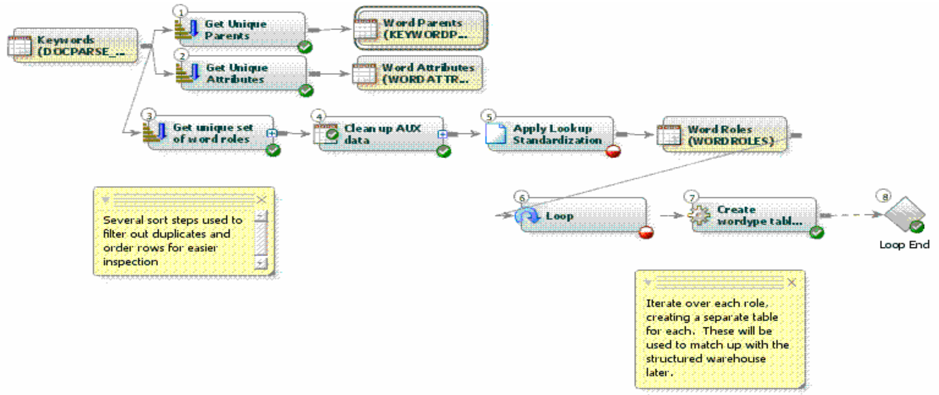
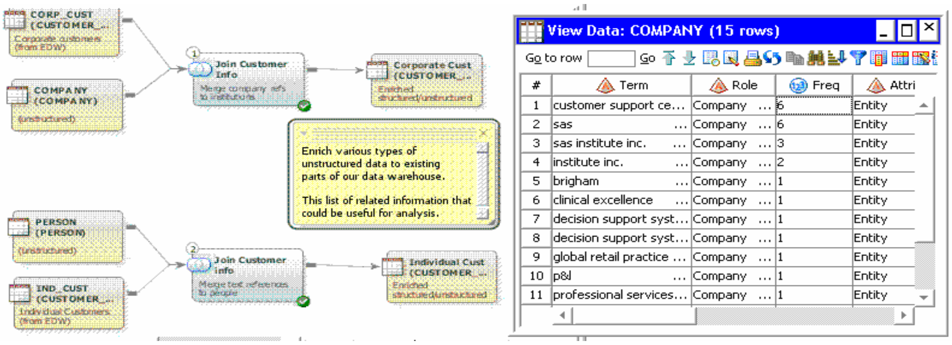

Keywords (DOCPARSE_...
1
Get Unique Parents
Word Parents (KEYWORDP...
2
Get Unique Attributes
Word Attributes (WORDATTR...
3
Get unique set of word roles
4
Clean up AUX data
5
Apply Lookup Standardization
Word Roles (WORDROLES)
6
Loop
7
Create wordype tabl...
8
Loop End
Several sort steps used to filter out duplicates and order rows for easier inspection
Iterate over each role, creating a separate table for each. These will be used to match up with the structured warehouse later.

Figure 16. Transforming Unstructured Textual Data To Enrich Existing Warehouse Data

The output of this process is a set of tables that contain data that is structured to enrich the existing warehouse data with information pulled from the text files.


CORP_CUST (CUSTOMER_...
Corporate customers (from EDW)
COMPANY (COMPANY)
(unstructured)
1
Join Customer Info
Merge company refs to institutions
Corporate Cust (CUSTOMER_...
Enriched structured/unstructured
PERSON (PERSON)
(unstructured)
IND_CUST (CUSTOMER_...
Individual Customers (from EDW)
2
Join Customer info
Merge text references to people
Individual Cust (CUSTOMER_...
Enriched structured/unstructured
Go to row
Go

Enrich various types of unstructured data to existing parts of our data warehouse.

This list of related information that could be useful for analysis.

View Data: COMPANY (15 rows)

#	Term	Role	Freq	Attri
1	customer support ce...	Company ...	6	Entity
2	sas ...	Company ...	6	Entity
3	sas institute inc. ...	Company ...	3	Entity
4	institute inc. ...	Company ...	2	Entity
5	brigham ...	Company ...	1	Entity
6	clinical excellence ...	Company ...	1	Entity
7	decision support syst...	Company ...	1	Entity
8	decision support syst...	Company ...	1	Entity
9	global retail practice ...	Company ...	1	Entity
10	p&l ...	Company ...	1	Entity
11	professional services...	Company ...	1	Entity

Figure 17. Final Result of the Data Integration Studio Job Showing Transformed Text Data Ready to Combine with the Existing Data Warehouse

This processing can be performance intensive because web logs and word counts can be large. Using the Data Integration Studio Loop transform to perform this processing in parallel on an SMP machine or on a Grid can significantly speed up the process.

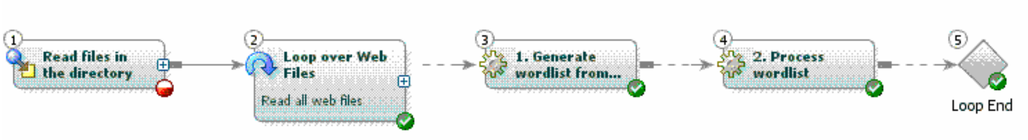


Figure 18. Performing the Text Transforms in Parallel Is Easy Using the Loop transform in SAS Data Integration Studio

PERFORMANCE MONITORING USING JMP

When monitoring the performance of grid nodes, you could find that you have to tune the environment to gain the best performance. Many variables can affect performance, for example, the number of machines in the grid, the amount of memory or cache per node, I/O capabilities, and the current workload. It is therefore easier to visualize the system performance rather than see individual node information or tables of performance numbers. SAS JMP is the ideal visualization tool for visualizing grid node performance, because JMP dynamically links statistics with graphics.

Using the census data sample scenario discussed earlier, JMP can help you see how various aspects of grid performance are affected by the range of tuning parameters that are available.

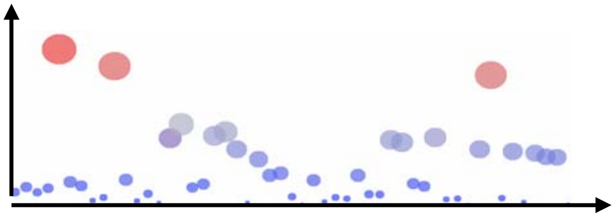


Figure 19. Performance Visualization Using JMP Shows That the Grid Performance Is Not Yet Optimized

Figure 19 is a JMP bubble plot that shows an untuned grid deployment using the sample census data. The X axis represents the state being processed, and the Y axis shows the duration of time that is required for the state. We can also use bubble size as an indicator for a value such as memory use, and use bubble color to represent another value such as I/O required to process that state. Looking at a collection of values makes tuning easier because you can identify trade-offs. What we want is the smallest circles, lowest on theY axis, and with the desired color. In the plot above the largest circles belong to California, Florida, and Texas because data from their larger populations requires more processing time.

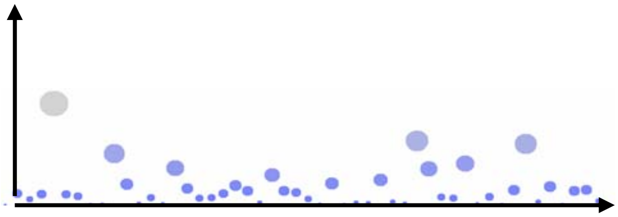


Figure 20. Well Tuned Grid Showing Optimal Performance Visualization

As tuning progresses, using the JMP software we can see as the network approaches the optimal configuration. Completed tuning results in the desired visual outcome: small circles (minimal memory used) that are lowest on the Y axis (smallest time requirement) for all states.

CONCLUSION

SAS parallelization capabilities have been continually enhanced with each release of SAS software. The introduction of SAS Grid Manager in SAS Version 9.1.3 marked a significant milestone in performance computing support in the SAS System. SAS Data Integration Studio makes it easy to leverage parallel and grid computing through customized transforms that enable you to parallelize your processes easily and efficiently. Data Integration Studio also supports all aspects of SAS Grid Computing so that you can select the technique that best matches your computing needs. Grid computing lends itself to those scenarios that require a significant amount of computational and I/O resources. The processing of unstructured data using SAS Data Integration Studio is an excellent example of one scenario that can benefit from the performance optimizations provided by SAS Grid processing. For performance tuning, SAS JMP software can help visualize the many variables required to optimize grid performance.

REFERENCES

Hunley, Eric, Gary Mehler, and Nancy Rausch. 2008. “A Whole New World: What’s New in SAS(R) Data Integration Studio 4.2.” *Proceedings of the SAS Global Forum 2008 Conference*. Cary, NC: SAS Institute Inc. Available at http://www2.sas.com/proceedings/forum2008.

SAS Institute Inc. 2007. “SAS Data Integration Studio 3.4 User’s Guide.” Cary, NC: SAS Institute Inc. Available at http://support.sas.com/documentation/onlinedoc/etls/index.html

Hewlett Packard. 2007. “SAS Benchmark Report for SAS® Grid Manager Running Linux on HP ProLiant BL460c Servers and HP StorageWorks.”
Available at http://h20219.www2.hp.com/ERC/downloads/4AA1-6058ENW.pdf

SAS Institute Inc. 2007. “SAS Demonstrates Superior Scalability with SAS® Grid Manager and Sun StorageTek QFS Shared File System.” Cary, NC: SAS Institute Inc.
Available at http://www.sas.com/partners/directory/sun/SASGridMgr0907.pdf

SAS Institute Inc. 2006. “SAS® Enterprise Data Integration Server Building a 4.2 Terabyte Star Schema Using SAS Grid Technology.” Cary, NC: SAS Institute Inc.

Introduction to Grid Computing. Scalability and Performance Community. SAS Institute Inc. 2007 Available at http://support.sas.com/rnd/scalability/grid/

SAS Grid Computing. Technologies/Enterprise Intelligence Platform. SAS Institute Inc. 2007 Available at http://www.sas.com/technologies/architecture/grid/index.html

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Cheryl Doninger
SAS Institute Inc.
Cary, NC 27513
Work Phone: (919) 677-8000
Fax: (919) 677-4444
E-mail: Cheryl.Doninger@sas.com
Web: www.sas.com

Gary Mehler
SAS Institute Inc.
Cary, NC 27513
Work Phone: (919) 677-8000
Fax: (919) 677-4444
E-mail: Gary.Mehler@sas.com
Web: www.sas.com

Nancy Rausch
SAS Institute Inc.
Cary, NC 27513
Work Phone: (919) 677-8000
Fax: (919) 677-4444
E-mail: Nancy.Rausch@sas.com
Web: www.sas.com