ABSTRACT
The Data Integration Monitor provides a SAS® data integration (DI) site a browser-based front-end to monitor the state and the progress of SAS DI batch processes in real time, quickly discover and solve SAS DI batch problems, and discover historical trends and potential bottlenecks in SAS DI batch processes. In addition, it provides a facility for gathering audit statistics for data quality and data governance purposes.

INTRODUCTION
The DI Monitor provides a SAS Data Integration (DI) Administrator a browser based front-end to monitor state and progress of SAS DI batch processes in real-time, quickly discover and solve SAS DI batch problems, and discover historical trends and potential bottlenecks in SAS DI batch processes. In addition it provides a facility for gathering audit statistics for data quality and data governance purposes.

This paper discusses the challenges that face a SAS Administrator when responsible for SAS DI batch jobs and flows, and how the DI Job Monitor can help.

This paper should be of interest to SAS Administrators concerned with monitoring status and performance of production DI jobs. The assumption is made that the reader is familiar with general data integration concepts and has a basic knowledge of databases, tables, and queries.

CHALLENGES FACING THE SAS DI ADMINISTRATOR
The SAS DI Administrator is responsible for the continuous operation of DI batch processes loading data into the data warehouse. He needs to be in control of the processes that surround the loading of the data warehouse to provide correct, timely, and validated data and information to data warehouse consumers. To be in control, he needs answers to questions like:

- Which jobs ran when?
- Did they complete successfully?
- How long did they run?
- How much data did they process?
- Why did some fail?
- Why did some run longer than normal?
- Why did some run shorter than normal?
- What are the trends in run-times, memory usage, and CPU times?
- What run-times and data volumes can be expected in 3 months time?
- Did all data delivered by the source system get loaded in the data warehouse?
- Did data delivered by the source system get loaded correctly in the data warehouse?

To answer these questions, the Administrator can gather and combine data from sources like:
- The job scheduling tool used to submit the batch jobs, like Platform's LSF
- The SAS program logs
- SAS Management Console

and turn that into information using his SAS knowledge, his business knowledge, and his common sense.

Since standard SAS Software does not provide tools for this task, the answers to the questions are most often obtained through manual ad-hoc querying in Enterprise Guide. While this is fine in essence, a tool that combines the
separate data sources, adds intelligence to it, and provides interactive analysis targeted to the DI Administrator’s tasks would be very welcome. Enter: the DI Monitor!

**ENTER: THE DI MONITOR**

The DI Monitor allows a SAS DI site to monitor state and progress of SAS DI batch processes in real-time, quickly discover and solve SAS DI batch problems, and discover historical trends and potential bottlenecks in SAS DI batch processes. In addition it provides a facility for gathering audit statistics for data quality and data governance purposes.

The DI Monitor provides three primary functions:

- **Real-time Monitoring**, including:
  - Flow, Job, and Job Step monitoring with real-time status and elapsed times.
  - Automatic discovery and analysis of abnormal performance
  - Quick pinpointing causes of program abends.

- **History Reporting**, for example:
  - Flow, Job, and Job Step elapsed times with box plots and trend graphs
  - Real Time, CPU Time and Memory Usage reporting

- **Gathering and reporting Audit Statistics**, like:
  - The number of observations processed in a job.
  - Hash totals to check completeness and correctness

The DI Monitor provides a web-based front-end to its users and has drill-down capabilities to allow for examining details. The screenshot below shows the initial screen of the real-time component in action.

![Figure 1 Screenshot of the real-time component.](image)

The following screenshot is taken from the initial screen of the history-reporting component:

![Figure 2 Screenshot of the history-reporting component.](image)
ARCHITECTURE

The DI Monitor collects data through two simple scripts that are called from the scheduler's script that starts off a SAS DI batch job. To enable this, the scheduler's job start script (SAS.BAT for LSF on Windows, sas.sh for LSF on Unix) is modified to include a call to the DI Monitor's pre-job just before the call to the actual SAS DI job, and a call to the DI Monitor's post-job right after the call to the SAS DI job. The collected data is stored in a data model and is made available to the DI Monitor web interface through SAS Stored Processes.
Figure 3 Architecture

JOB STATISTICS

The pre-job is a simple SAS program that inserts a record into a database table with the name of the DI job, the flow it is running in, the job's start date and time, and the location of the SAS log file, where the status of the job is set to 'RUNNING'. The post-job updates the previously inserted record for this job with the job's end date and time and with its return code, setting the status of the job to either 'COMPLETED' or 'FAILED', depending on the job's return code.

An example table is:

<table>
<thead>
<tr>
<th>RUN_ID</th>
<th>FLOW</th>
<th>JOB</th>
<th>STARTDTS</th>
<th>ENDDTS</th>
<th>RC</th>
<th>SASLOG</th>
<th>STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Load</td>
<td>Orders</td>
<td>05APR11:12:34</td>
<td>05APR11:11:41</td>
<td>0</td>
<td>C:\Temp\Extract1.log</td>
<td>COMPLETED</td>
</tr>
<tr>
<td>1</td>
<td>Load</td>
<td>Orders</td>
<td>05APR11:11:42</td>
<td></td>
<td></td>
<td>C:\Temp\Transform1.log</td>
<td>RUNNING</td>
</tr>
</tbody>
</table>

This information is gathered on the flow, job and job step level.

PERFORMANCE STATISTICS

In addition to recording job completion statistics, the post-job has the option of parsing the SAS log and storing relevant information in a database table. This information includes the statistics that are generated by the FULLSTIMER option. For SAS91, the SAS-provided %LOGPARSE macro is used to parse the SAS log (see http://support.sas.com/kb/34/301.html). For SAS92 the ARM log statistics are used. Because parsing the SAS log file can be resource intensive, this option can be switched on and off in the dimon.ini file.

1 This is a simplified lay-out. The actual data model in which this information is stored holds more tables and is normalized.
AUDIT STATISTICS

Another option in the post-job is the execution of so-called Audit Queries that produce Audit Statistics. Audit Queries are user-written SQL queries stored in a database table and executed when a job completes. The Audit Query results can be used for data quality and data governance purposes.

An example of records in the Audit Query table is the following:

<table>
<thead>
<tr>
<th>ID</th>
<th>QUERY_CODE</th>
<th>QUERY_DESC</th>
<th>QUERY_RESULT_TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SELECT COUNT(*) FROM DWH.DIM_PRODUCT</td>
<td>A record count of the Products dimension table</td>
<td>NUMERIC</td>
</tr>
<tr>
<td>2</td>
<td>SELECT SUM(ORDER_NR) FROM STAGING.CUSTOMERS</td>
<td>Order numbers hash total on the Orders staging table</td>
<td>NUMERIC</td>
</tr>
</tbody>
</table>

A control table holds the information on which query is to be executed after which jobs in which flows.

Audit Query results are inserted into the Audit Query Results table, for instance:

<table>
<thead>
<tr>
<th>RUN_ID</th>
<th>FLOW</th>
<th>JOB</th>
<th>AUDIT_QUERY_ID</th>
<th>RESULT_NUM</th>
<th>RESULT_CHAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Load Products</td>
<td>Load</td>
<td>1</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Load Orders</td>
<td>Extract</td>
<td>2</td>
<td>30010</td>
<td></td>
</tr>
</tbody>
</table>

THE WEB APPLICATION

The web application has two components:
- Real-time monitoring
- History reporting

These components are discussed below.

REAL-TIME MONITORING

In the figure below, job statistics on the flow level are shown.

Figure 4 Job Flow Statistics
Some flows have an exclamation mark in front of their elapsed time statistic, which indicates that the elapsed time deviates significantly from the last 60 days average. When clicking on such an elapsed time, an automatic analysis is made of deviations from normal for:

- Elapsed times
- Counts of the number of records processed. Deviations may indicate that source files are not complete or process steps returned a data set that is significantly larger or smaller than normal.
- Ratio of real-time vs. CPU time, which can be used as an indicator for system load.

The following screen results from clicking on the elapsed time link for the Load Orders flow.

![Figure 5 Flow details](image)

Drilling down on the flow name “Load Orders” takes you to the job overview of the flow:

![Figure 6 Job overview of selected flow.](image)
Drilling down on the job takes you to the job step overview of the selected job:

![Job Step Performance Statistics for Job: Integrate - Orders]

**Figure 7** Job step overview of selected job.

Drilling down on the job step takes you to the SAS log of the selected job step:

```sas
/*-----------------------------*/
56  * Step:   SAS Extract        ASUSHEGF.EW00033A ^
57  * Transform: Extract         ^
58  * Description:              ^
59  * Source Table: JOB_RUN_STS  ^
60  * Target Table: Extract_Tar ^
61  * Primary Key:              ^
62  *                                        ^
63  /*                          ^
64  %let SYSLIB = &sysdata(JOB_RUN_STS);%
65  %let transformID = &sysdata(ASUSHEGF.EW00033A);
66  %let trans_fc = 0;
67  proc datasets lib = work.noshow newtype = (date viable);
68     delete W0047TF; quit;
```

**Figure 8** SAS log of selected job step
HISTORY REPORTING

The initial screen of the history-reporting component of the DI monitor is shown below:

![Figure 9 History reporting component of the DI Monitor](image)

It shows statistics for the last 90 days:
- A horizontal bar of the mean elapsed time per flow
- A horizontal box plot of the elapsed time per flow
- A trend graph of the elapsed time per flow.

Clicking the trend graph gives further details on the trend, including a regression equation:

![Figure 10 Details for selected trend graph](image)

Regression Equation: ELAPSED_TIME = -1929.100 + 0.11063*RUN_DATE.
Clicking a flow from the flow history overview takes you to the history overview of the jobs in the selected flow:

![Figure 11 History of jobs in selected flow](image)

Drilling down on a job takes you to the history of job steps in the selected job:

![Figure 12 History of job steps in selected job](image)
AVAILABILITY

The DI Monitor is provided as open source software. Please contact the author at bheinsius@eom.nl to obtain it.

CONCLUSION

The DI Monitor provides a SAS Data Integration (DI) Administrator a valuable tool for monitoring state and progress of SAS DI batch processes in real-time, quickly discover and solve SAS DI batch problems, and find historical trends and potential future bottlenecks in SAS DI batch processes. In addition it provides a facility for gathering audit statistics for data quality and data governance purposes.

The DI Monitor is provided as open source software. Please contact the author at bheinsius@eom.nl to obtain it.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Bart Heinsius  
Enterprise: EOM Data Solutions BV  
Address: Splijtbakweg 117  
City, State ZIP: 1333 HJ Almere, the Netherlands  
Work Phone: +31 85 877 8984  
Fax: +31 36 845 0224  
E-mail: bheinsius@eom.nl  
Web: http://www.eom.nl

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.