Paper 138-2011

Data Integration Monitor

Bart Heinsius, EOM Data Solutions BV, Almere, Netherlands

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!

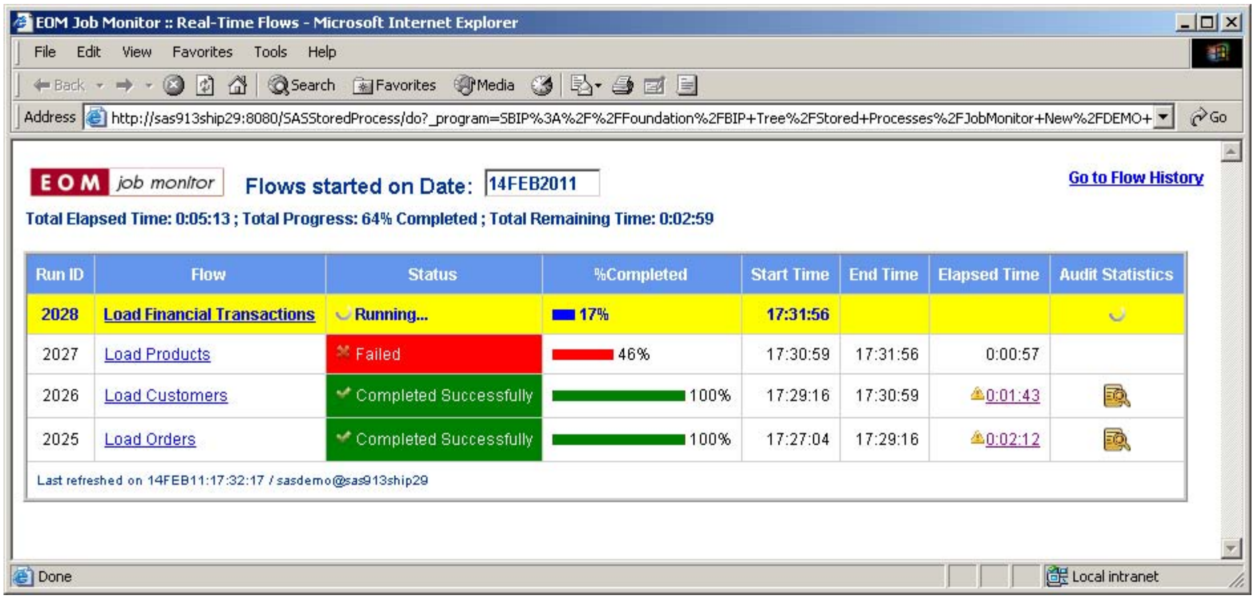
ENTER: THE DI MONITOR

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.


EOM Job Monitor :: Real-Time Flows - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go
EOM job monitor
Go to Flow History
Done
Local intranet

Address: http://sas913ship29:8080/SASStoredProcess/do?_program=SBIP%3A%2F%2FFoundation%2FBIP+Tree%2FStored+Processes%2FJobMonitor+New%2FDEMO+

Flows started on Date: 14FEB2011

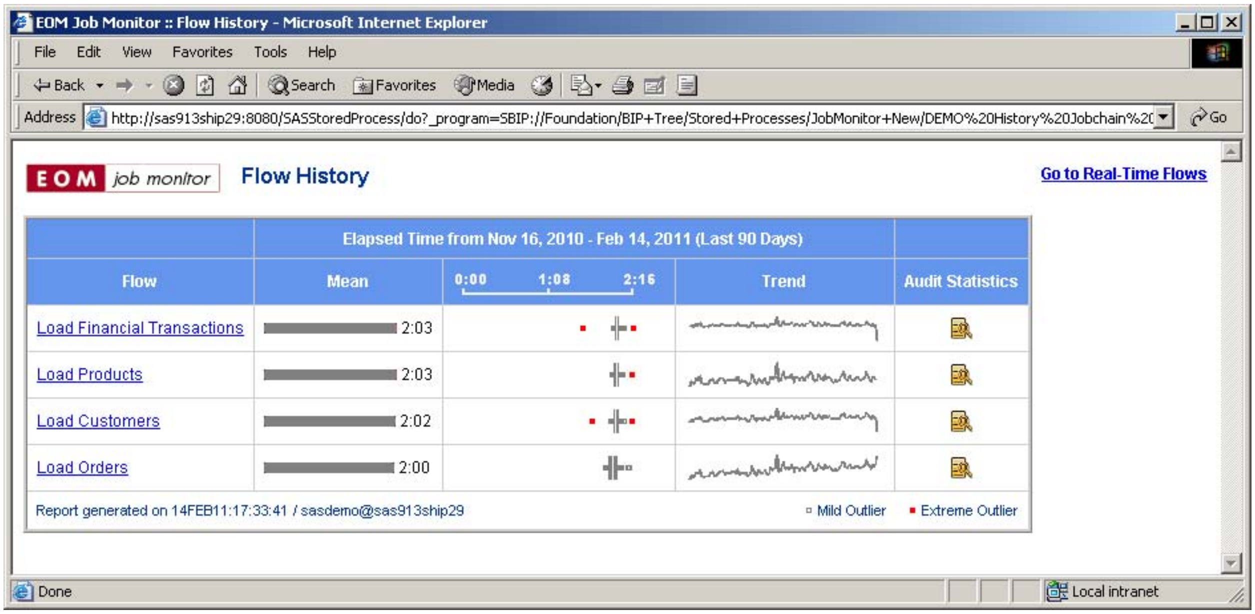
Total Elapsed Time: 0:05:13 ; Total Progress: 64% Completed ; Total Remaining Time: 0:02:59

Run ID	Flow	Status	%Completed	Start Time	End Time	Elapsed Time	Audit Statistics
2028	Load Financial Transactions	Running...	17%	17:31:56			
2027	Load Products	✗ Failed	46%	17:30:59	17:31:56	0:00:57	
2026	Load Customers	✓ Completed Successfully	100%	17:29:16	17:30:59	⚠0:01:43	
2025	Load Orders	✓ Completed Successfully	100%	17:27:04	17:29:16	⚠0:02:12	

Last refreshed on 14FEB11:17:32:17 / sasdemo@sas913ship29

Figure 1 Screenshot of the real-time component.

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


EOM Job Monitor :: Flow History - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go
Done
Local intranet
EOM job monitor

http://sas913ship29:8080/SASStoredProcess/do?_program=SBIP://Foundation/BIP+Tree/Stored+Processes/JobMonitor+New/DEMO%20History%20Jobchain%2(

Flow History

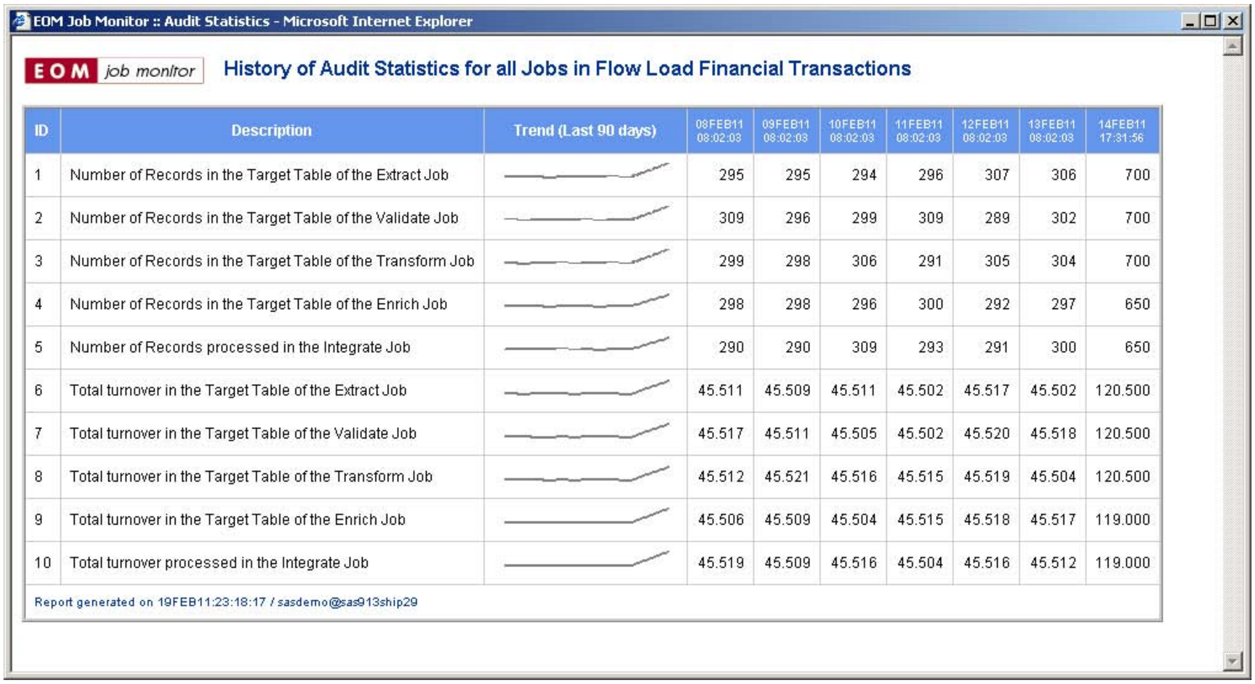
Go to Real-Time Flows

	Elapsed Time from Nov 16, 2010 - Feb 14, 2011 (Last 90 Days)			
Flow	Mean	0:00 1:08 2:16	Trend	Audit Statistics
Load Financial Transactions	2:03			
Load Products	2:03			
Load Customers	2:02			
Load Orders	2:00			

Report generated on 14FEB11:17:33:41 / sasdemo@sas913ship29 ▫ Mild Outlier ▪ Extreme Outlier

Figure 2 Screenshot of the history-reporting component.

The following screenshot is taken from an audit statistics history overview.


EOM Job Monitor :: Audit Statistics - Microsoft Internet Explorer
EOM job monitor

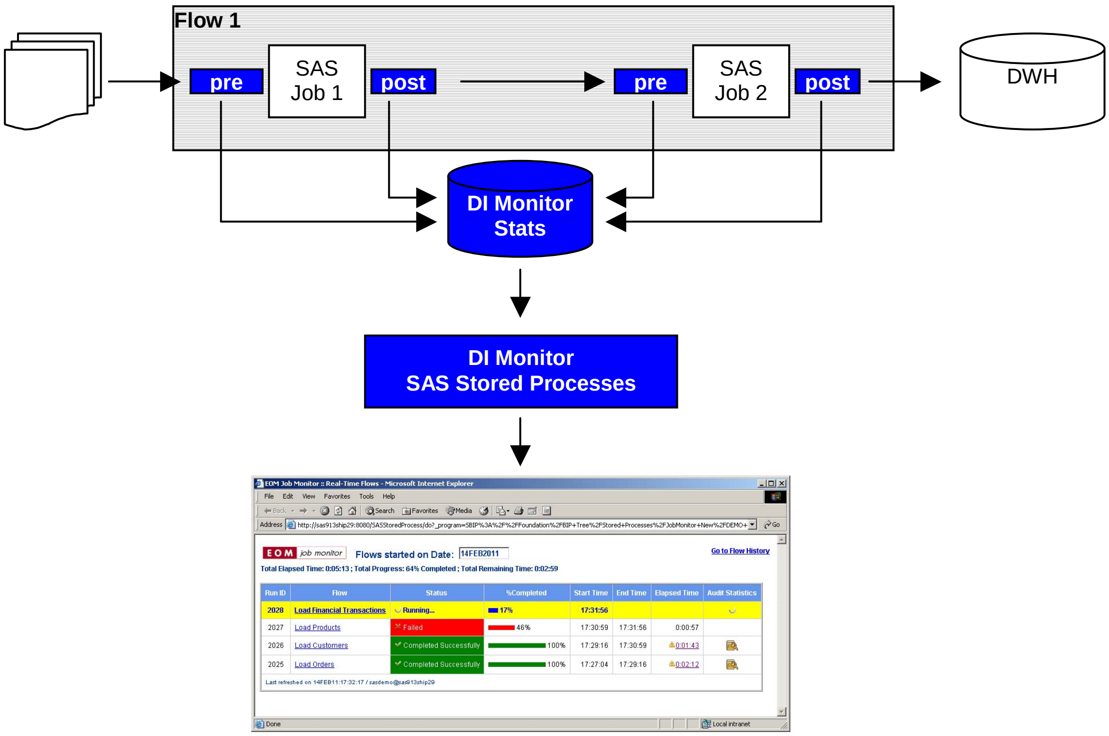
History of Audit Statistics for all Jobs in Flow Load Financial Transactions

ID	Description	Trend (Last 90 days)	08FEB11 08:02:03	09FEB11 08:02:03	10FEB11 08:02:03	11FEB11 08:02:03	12FEB11 08:02:03	13FEB11 08:02:03	14FEB11 17:31:56
1	Number of Records in the Target Table of the Extract Job		295	295	294	296	307	306	700
2	Number of Records in the Target Table of the Validate Job		309	296	299	309	289	302	700
3	Number of Records in the Target Table of the Transform Job		299	298	306	291	305	304	700
4	Number of Records in the Target Table of the Enrich Job		298	298	296	300	292	297	650
5	Number of Records processed in the Integrate Job		290	290	309	293	291	300	650
6	Total turnover in the Target Table of the Extract Job		45.511	45.509	45.511	45.502	45.517	45.502	120.500
7	Total turnover in the Target Table of the Validate Job		45.517	45.511	45.505	45.502	45.520	45.518	120.500
8	Total turnover in the Target Table of the Transform Job		45.512	45.521	45.516	45.515	45.519	45.504	120.500
9	Total turnover in the Target Table of the Enrich Job		45.506	45.509	45.504	45.515	45.518	45.517	119.000
10	Total turnover processed in the Integrate Job		45.519	45.509	45.516	45.504	45.516	45.512	119.000

Report generated on 19FEB11:23:18:17 / sasdemo@sas913ship29

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.


Flow 1
pre
SAS Job 1
post
pre
SAS Job 2
post
DWH
DI Monitor Stats
DI Monitor SAS Stored Processes
EOM Job Monitor :: Real-Time Flows - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://sas913ship29:8080/SASStoredProcess/do?_program=SBIP%3A%2F%2FFoundation%2FBIP+Tree%2FStored+Processes%2FJobMonitor+New%2FDEMO+
Go
EOM job monitor
Go to Flow History
Done
Local intranet

Flows started on Date: 14FEB2011

Total Elapsed Time: 0:05:13 ; Total Progress: 64% Completed ; Total Remaining Time: 0:02:59

Run ID	Flow	Status	%Completed	Start Time	End Time	Elapsed Time	Audit Statistics
2028	Load Financial Transactions	Running...	17%	17:31:56			
2027	Load Products	Failed	46%	17:30:59	17:31:56	0:00:57	
2026	Load Customers	Completed Successfully	100%	17:29:16	17:30:59	0:01:43	
2025	Load Orders	Completed Successfully	100%	17:27:04	17:29:16	0:02:12	

Last refreshed on 14FEB11:17:32:17 / sasdemo@sas913ship29

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[1]:

RUN_ID	FLOW	JOB	STARTDTS	ENDDTS	RC	SASLOG	STATUS
1	Load Orders	Extract	05APR11:12:34	05APR11:11:41	0	C:\Temp\Extract1.log	COMPLETED
1	Load Orders	Transform	0 5APR11:11:42			C:\Temp\Transform1.log	RUNNING

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:

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

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:

RUN_ID	FLOW	JOB	AUDIT_QUERY_ ID	RESULT_NUM	RESULT_CHAR
1	Load Products	Load	1	200	
2	Load Orders	Extract	2	30010	

THE WEB APPLICATION

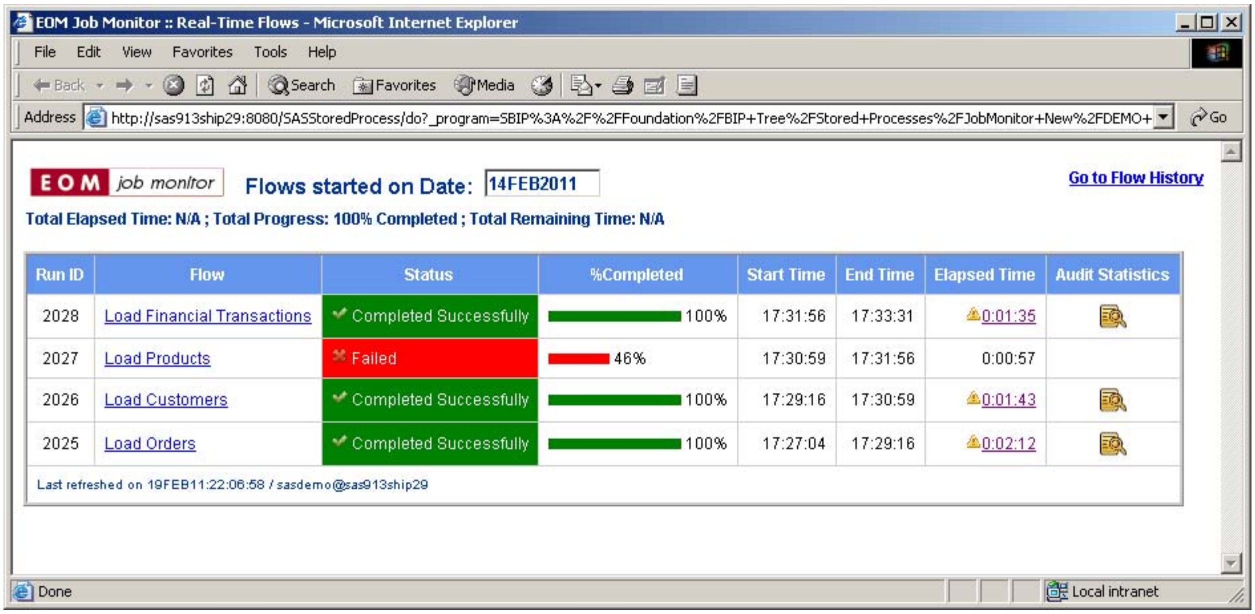
The web application has two components:

- 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.



http://sas913ship29:8080/SASStoredProcess/do?_program=SBIP%3A%2F%2FFoundation%2FBIP+Tree%2FStored+Processes%2FJobMonitor+New%2FDEMO+

Flows started on Date: 14FEB2011

Total Elapsed Time: N/A ; Total Progress: 100% Completed ; Total Remaining Time: N/A

Run ID	Flow	Status	%Completed	Start Time	End Time	Elapsed Time	Audit Statistics
2028	Load Financial Transactions	✓ Completed Successfully	100%	17:31:56	17:33:31	⚠0:01:35	🔍
2027	Load Products	✗ Failed	46%	17:30:59	17:31:56	0:00:57	
2026	Load Customers	✓ Completed Successfully	100%	17:29:16	17:30:59	⚠0:01:43	🔍
2025	Load Orders	✓ Completed Successfully	100%	17:27:04	17:29:16	⚠0:02:12	🔍

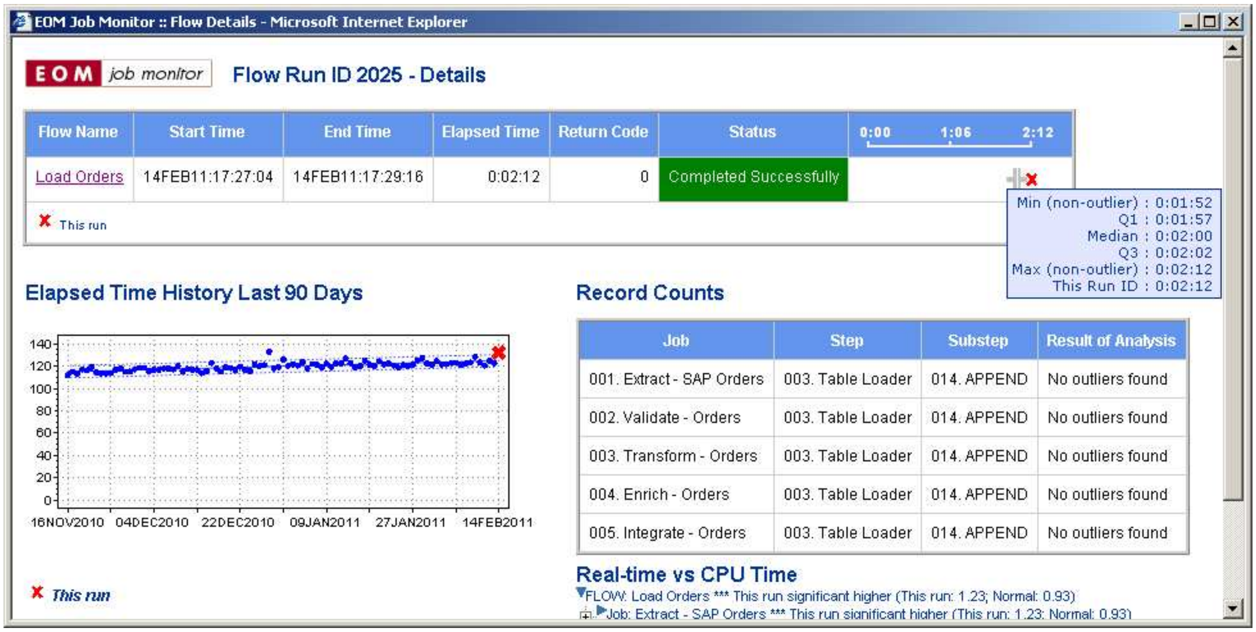
Last refreshed on 19FEB11:22:06:58 / sasdemo@sas913ship29

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 that 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.


EOM Job Monitor :: Flow Details - Microsoft Internet Explorer
EOM job monitor

Flow Run ID 2025 - Details

Flow Name	Start Time	End Time	Elapsed Time	Return Code	Status	0:00 1:06 2:12
Load Orders	14FEB11:17:27:04	14FEB11:17:29:16	0:02:12	0	Completed Successfully	

✗ This run

Min (non-outlier) : 0:01:52
Q1 : 0:01:57
Median : 0:02:00
Q3 : 0:02:02
Max (non-outlier) : 0:02:12
This Run ID : 0:02:12

Elapsed Time History Last 90 Days

Record Counts

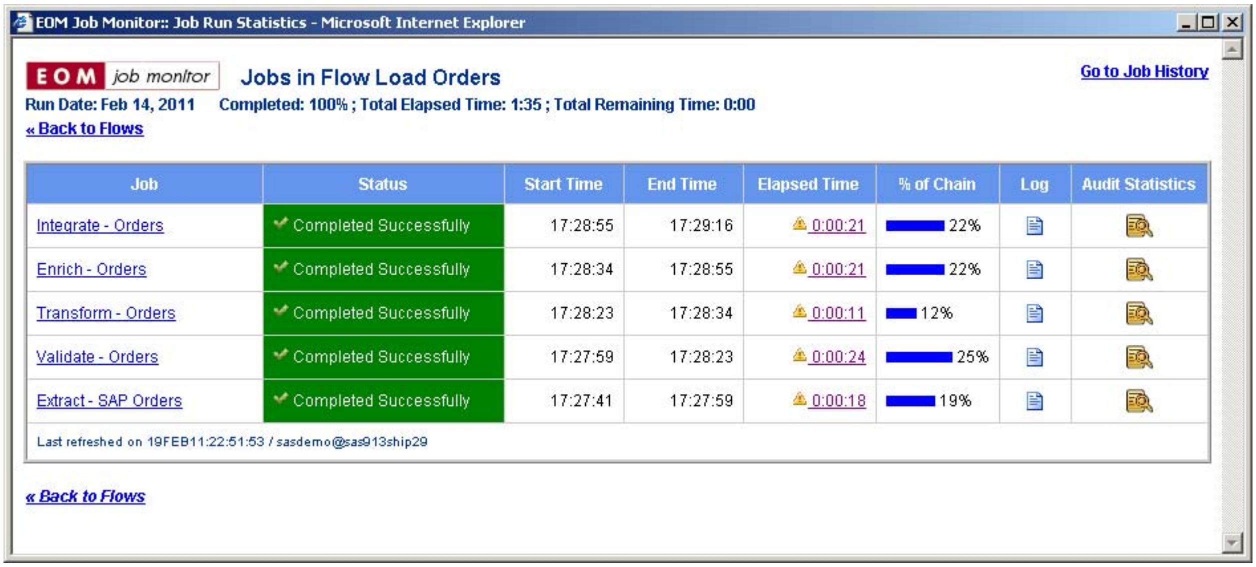
Job	Step	Substep	Result of Analysis
001. Extract - SAP Orders	003. Table Loader	014. APPEND	No outliers found
002. Validate - Orders	003. Table Loader	014. APPEND	No outliers found
003. Transform - Orders	003. Table Loader	014. APPEND	No outliers found
004. Enrich - Orders	003. Table Loader	014. APPEND	No outliers found
005. Integrate - Orders	003. Table Loader	014. APPEND	No outliers found

Real-time vs CPU Time

FLOW: Load Orders *** This run significant higher (This run: 1.23; Normal: 0.93)
Job: Extract - SAP Orders *** This run significant higher (This run: 1.23; Normal: 0.93)

Figure 5 Flow details

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


EOM Job Monitor:: Job Run Statistics - Microsoft Internet Explorer
EOM job monitor

Jobs in Flow Load Orders

Go to Job History

Run Date: Feb 14, 2011 Completed: 100% ; Total Elapsed Time: 1:35 ; Total Remaining Time: 0:00

« Back to Flows

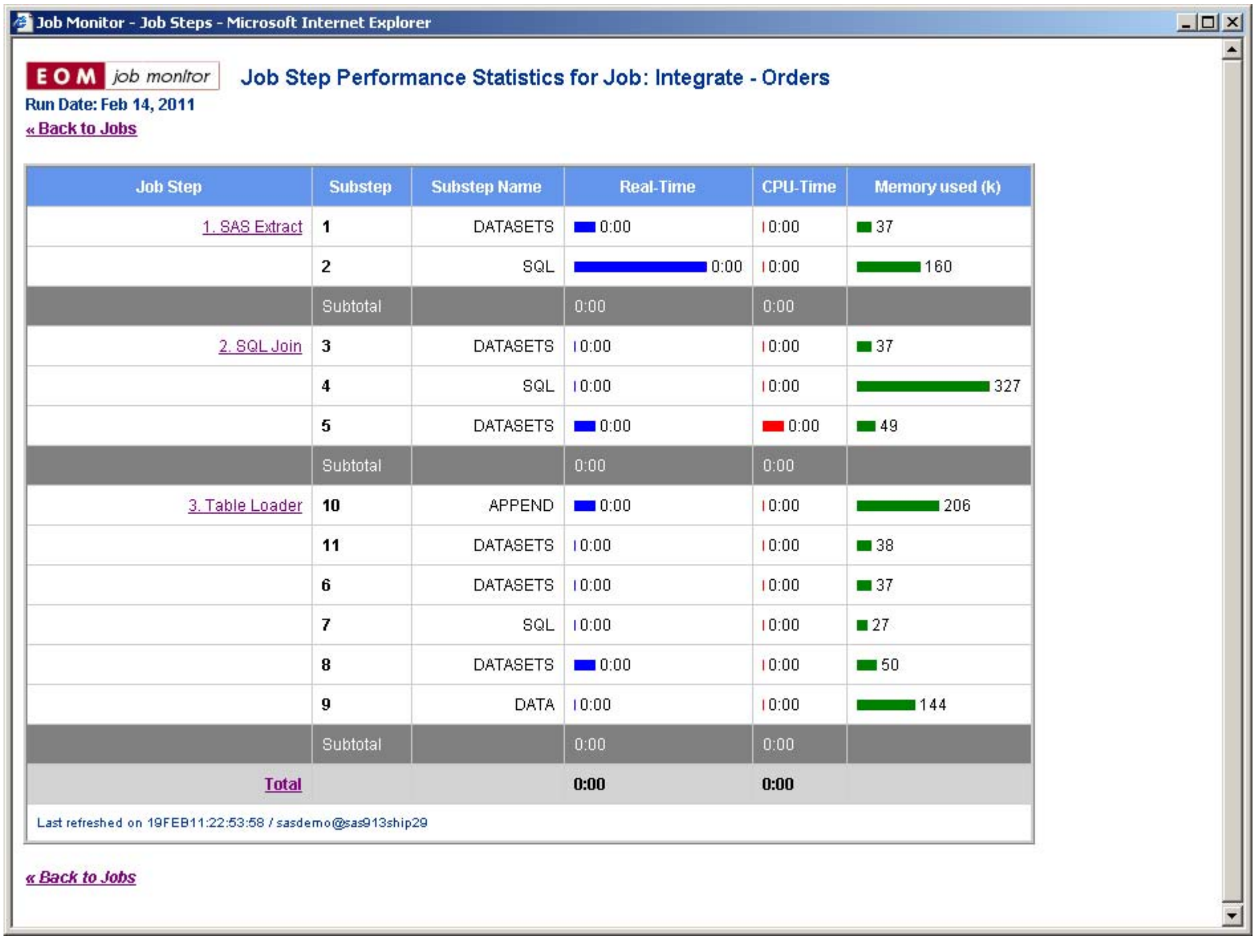
Job	Status	Start Time	End Time	Elapsed Time	% of Chain	Log	Audit Statistics
Integrate - Orders	✓ Completed Successfully	17:28:55	17:29:16	⚠ 0:00:21	22%		
Enrich - Orders	✓ Completed Successfully	17:28:34	17:28:55	⚠ 0:00:21	22%		
Transform - Orders	✓ Completed Successfully	17:28:23	17:28:34	⚠ 0:00:11	12%		
Validate - Orders	✓ Completed Successfully	17:27:59	17:28:23	⚠ 0:00:24	25%		
Extract - SAP Orders	✓ Completed Successfully	17:27:41	17:27:59	⚠ 0:00:18	19%		

Last refreshed on 19FEB11:22:51:53 / sasdemo@sas913ship29

« Back to Flows

Figure 6 Job overview of selected flow.

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



Job Monitor - Job Steps - Microsoft Internet Explorer

EOM job monitor

Job Step Performance Statistics for Job: Integrate - Orders

Run Date: Feb 14, 2011

« Back to Jobs

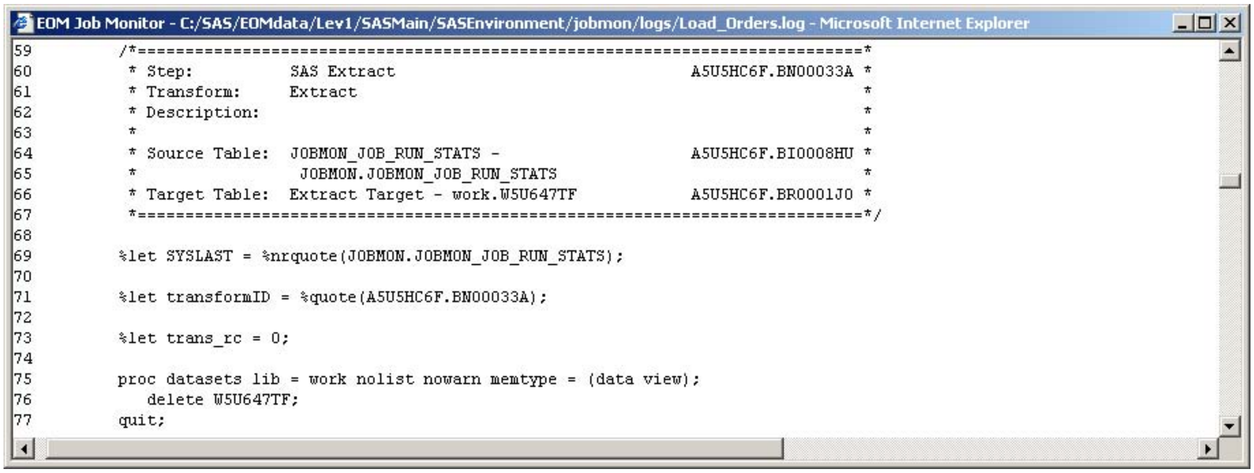
Job Step	Substep	Substep Name	Real-Time	CPU-Time	Memory used (k)
1. SAS Extract	1	DATASETS	0:00	0:00	37
	2	SQL	0:00	0:00	160
	Subtotal		0:00	0:00	
2. SQL Join	3	DATASETS	0:00	0:00	37
	4	SQL	0:00	0:00	327
	5	DATASETS	0:00	0:00	49
	Subtotal		0:00	0:00	
3. Table Loader	10	APPEND	0:00	0:00	206
	11	DATASETS	0:00	0:00	38
	6	DATASETS	0:00	0:00	37
	7	SQL	0:00	0:00	27
	8	DATASETS	0:00	0:00	50
	9	DATA	0:00	0:00	144
	Subtotal		0:00	0:00	
Total			0:00	0:00	

Last refreshed on 19FEB11:22:53:58 / sasdemo@sas913ship29

« Back to Jobs

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:



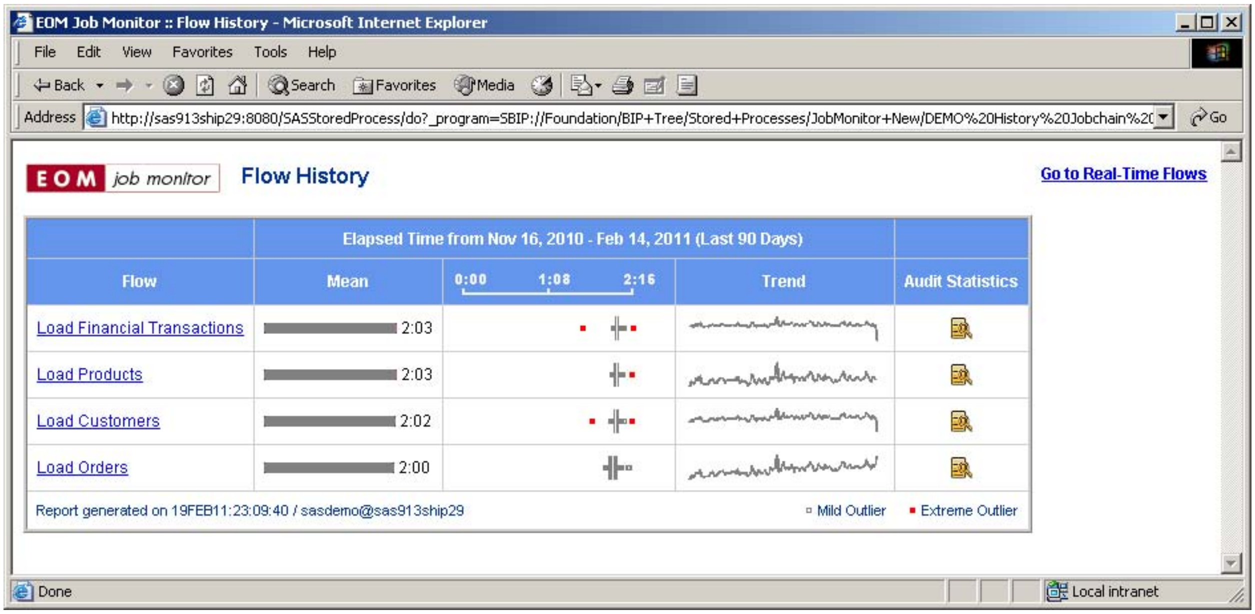
EOM Job Monitor - C:/SAS/EOMdata/Lev1/SASMain/SASEnvironment/jobmon/logs/Load_Orders.log - Microsoft Internet Explorer

```
/*==========================================================================*
 * Step:          SAS Extract                            A5U5HC6F.BN00033A *
 * Transform:     Extract                                                  *
 * Description:                                                            *
 *                                                                         *
 * Source Table:  JOBMON_JOB_RUN_STATS -                 A5U5HC6F.BI0008HU *
 *                 JOBMON.JOBMON_JOB_RUN_STATS                             *
 * Target Table:  Extract Target - work.W5U647TF         A5U5HC6F.BR0001J0 *
 *==========================================================================*/

%let SYSLAST = %nrquote(JOBMON.JOBMON_JOB_RUN_STATS);

%let transformID = %quote(A5U5HC6F.BN00033A);

%let trans_rc = 0;

proc datasets lib = work nolist nowarn memtype = (data view);
   delete W5U647TF;
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:


EOM Job Monitor :: Flow History - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go
EOM job monitor
Done
Local intranet

http://sas913ship29:8080/SASStoredProcess/do?_program=SBIP://Foundation/BIP+Tree/Stored+Processes/JobMonitor+New/DEMO%20History%20Jobchain%20

Flow History

Go to Real-Time Flows

	Elapsed Time from Nov 16, 2010 - Feb 14, 2011 (Last 90 Days)			
Flow	Mean	0:00 1:08 2:16	Trend	Audit Statistics
Load Financial Transactions	2:03			
Load Products	2:03			
Load Customers	2:02			
Load Orders	2:00			

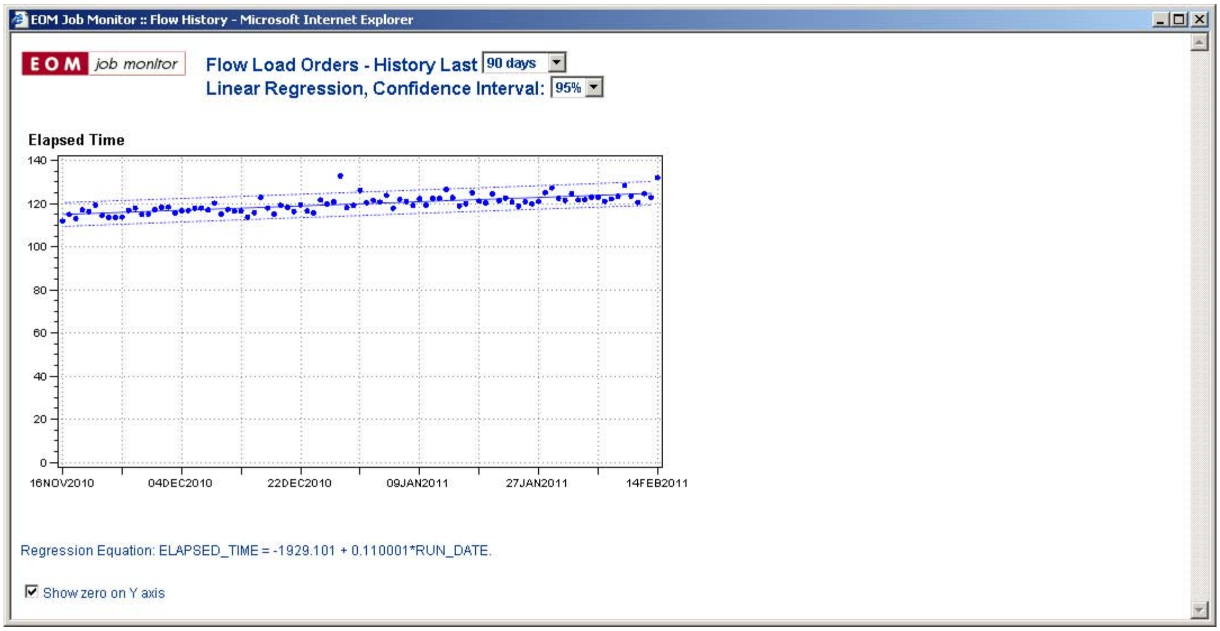
Report generated on 19FEB11:23:09:40 / sasdemo@sas913ship29 ▫ Mild Outlier ▪ Extreme Outlier

Figure 9 History reporting component of the DI Monitor

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:



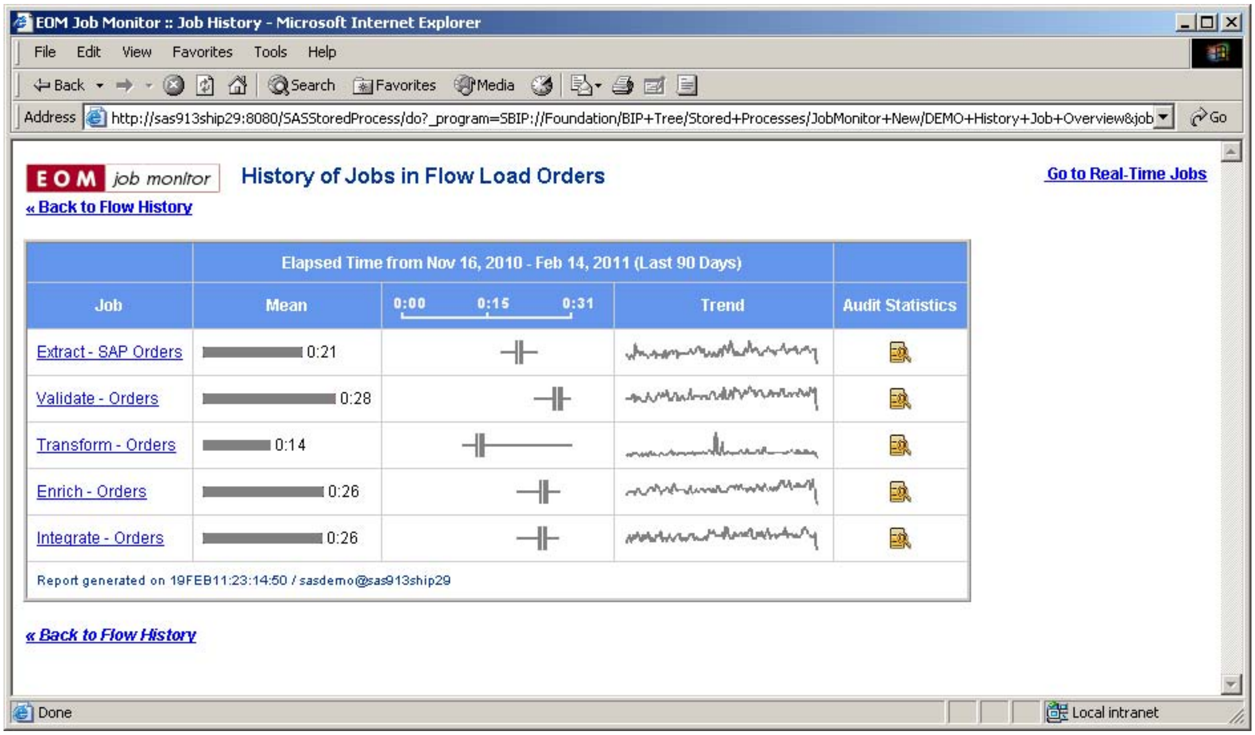
Flow Load Orders - History Last 90 days
Linear Regression, Confidence Interval: 95%

Regression Equation: ELAPSED_TIME = -1929.101 + 0.110001*RUN_DATE.

☑ Show zero on Y axis

Figure 10 Details for selected trend graph

Clicking a flow from the flow history overview takes you to the history overview of the jobs in the selected flow:


EOM Job Monitor :: Job History - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go
EOM job monitor
Done
Local intranet

Address: http://sas913ship29:8080/SASStoredProcess/do?_program=SBIP://Foundation/BIP+Tree/Stored+Processes/JobMonitor+New/DEMO+History+Job+Overview&job

History of Jobs in Flow Load Orders

« Back to Flow History

Go to Real-Time Jobs

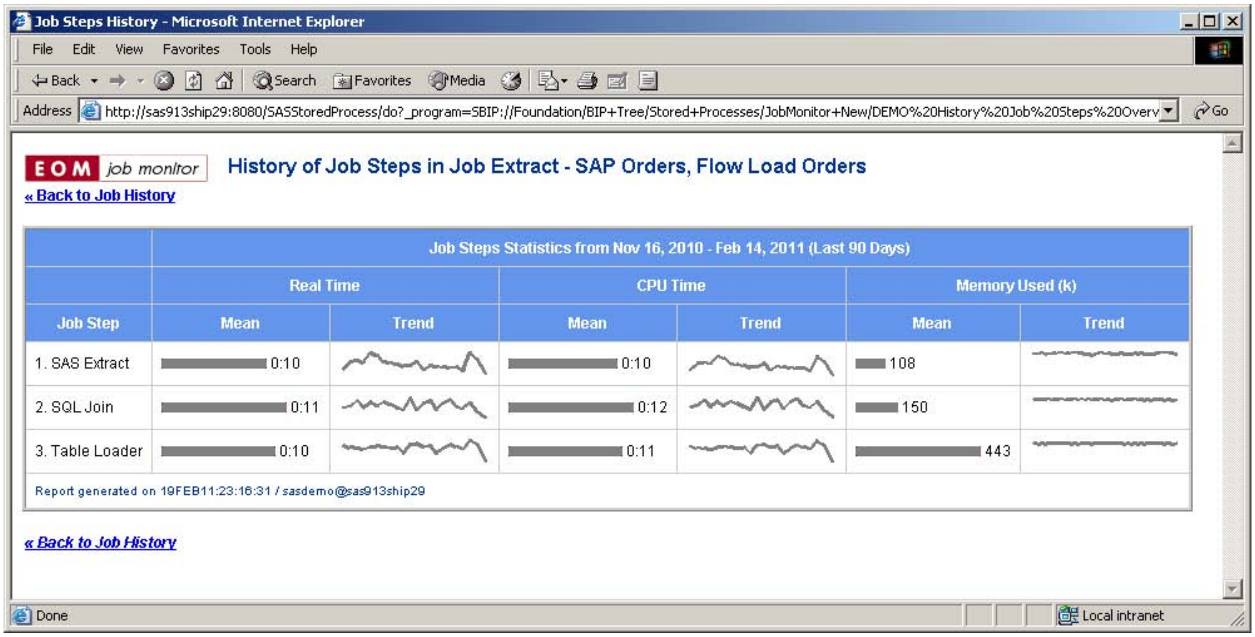
	Elapsed Time from Nov 16, 2010 - Feb 14, 2011 (Last 90 Days)			
Job	Mean	0:00 0:15 0:31	Trend	Audit Statistics
Extract - SAP Orders	0:21			
Validate - Orders	0:28			
Transform - Orders	0:14			
Enrich - Orders	0:26			
Integrate - Orders	0:26			

Report generated on 19FEB11:23:14:50 / sasdemo@sas913ship29

« Back to Flow History

Figure 11 History of jobs in selected flow

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


Job Steps History - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go
EOM job monitor
Done
Local intranet

Address: http://sas913ship29:8080/SASStoredProcess/do?_program=SBIP://Foundation/BIP+Tree/Stored+Processes/JobMonitor+New/DEMO%20History%20Job%20Steps%20Overv

History of Job Steps in Job Extract - SAP Orders, Flow Load Orders

« Back to Job History

	Job Steps Statistics from Nov 16, 2010 - Feb 14, 2011 (Last 90 Days)					
	Real Time		CPU Time		Memory Used (k)	
Job Step	Mean	Trend	Mean	Trend	Mean	Trend
1. SAS Extract	0:10		0:10		108	
2. SQL Join	0:11		0:12		150	
3. Table Loader	0:10		0:11		443	

Report generated on 19FEB11:23:16:31 / sasdemo@sas913ship29

« Back to Job History

Figure 12 History of job steps in selected job

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