

Paper 327-2010

What's New in DataFlux® dfPower® Studio and DataFlux Integration Server Software?

Wilbram Hazejager, DataFlux Corporation, Heidelberg, Germany

ABSTRACT

The latest release of DataFlux® software contains many new features that help both in the operational data quality area and in the warehousing and analytics areas. Of particular interest to SAS® users will be the improved management of SAS data sources and targets. Other highlights include an improved user interface, performance enhancements, new data exploration features, and seamless, enhanced support for XML data sources, including support for HIPAA, X12, HL7, and others. This paper will discuss the new and improved features that DataFlux has delivered with its latest product set.

INTRODUCTION

DataFlux, a wholly owned subsidiary of SAS, provides end-to-end data management solutions that enable companies to effectively analyze, improve, and control their data. This paper will concentrate on DataFlux dfPower Studio and DataFlux Integration Server software.

dfPower Studio is a powerful, easy-to-use suite of data cleansing and data integration software applications. dfPower Studio connects to virtually any data source, and any dfPower Studio user in any department can use dfPower Studio to profile, cleanse, integrate, enrich, monitor, and otherwise improve data quality throughout the enterprise. dfPower Architect, an innovative job flow builder, enables users to build complex management workflows quickly and logically.

This capability allows front-line staff—not IT or development resources—to discover and address data problems, to merge customer and prospect databases, to verify and complete address information, to transform and standardize product codes, and to perform other data management processes that are required by your organization.

DataFlux Integration Server supports the ability to run batch dfPower Studio jobs in a client/server environment. DataFlux Integration Server is available in three editions: the DataFlux Standard Integration Server, DataFlux Enterprise Integration Server, and DataFlux Integration Server for SAS.

DataFlux Standard Integration Server enables any dfPower Studio client user to offload batch dfPower Profile and dfPower Architect jobs to a more scalable server environment.

DataFlux Integration Server for SAS additionally enables allows SAS programs to invoke DataFlux batch jobs from a SAS program and enables allows SAS programs to call a business service that was designed in the dfPower Studio client environment.

DataFlux Enterprise Integration Server has added the ability to allow the calling of business services that were designed in the dfPower Studio client environment, or to invoke batch jobs using service-oriented architecture (SOA) from any third-party application.

This paper highlights the new features and capabilities in the latest release of dfPower Studio and DataFlux Integration Server. All of the features discussed here are available in DataFlux release 8.1 or newer, unless indicated otherwise.

IMPROVED USER EXPERIENCE

With the introduction of DataFlux software release 8.0, the main user interface, called dfPower Studio Navigator, has been significantly enhanced to provide easier access to the different functionalities and to make it easier to manage all the objects that have been created by the user. DataFlux now uses Unicode, which enables all character sets to be processed and displayed in the user interface.

Furthermore, Job Editing functionalities have been enhanced to increase productivity. Key features include:

- Copy/Paste support for node(s)
- Auto-Arrange nodes of nodes on a page
- Find node on page
- Ability to control where a note is inserted
- Ability to move multiple fields in the node output list

- Undo/Redo support (release 8.2 and above)

Managing access to SAS data sets has been improved by providing a central location to define directories that contain SAS data sets.

NEW NAVIGATOR

The main interface, called dfPower Studio Navigator, now supports toolbars, a navigation tree that gives access to all objects defined by the user. When the user double-clicks on a specific item in the navigation tree, a panel opens on the right side of the screen, providing a detailed overview of all the objects that are part of that category. In the screenshot below, one specific Management Resource was opened to provide an overview of all reports and jobs that are defined in that Management Resource. Within a Management Resource, user-defined groups are now supported to better organize and manage jobs and reports. Also, multiple Management Resources can be defined (for example, to keep jobs and reports separated by project).

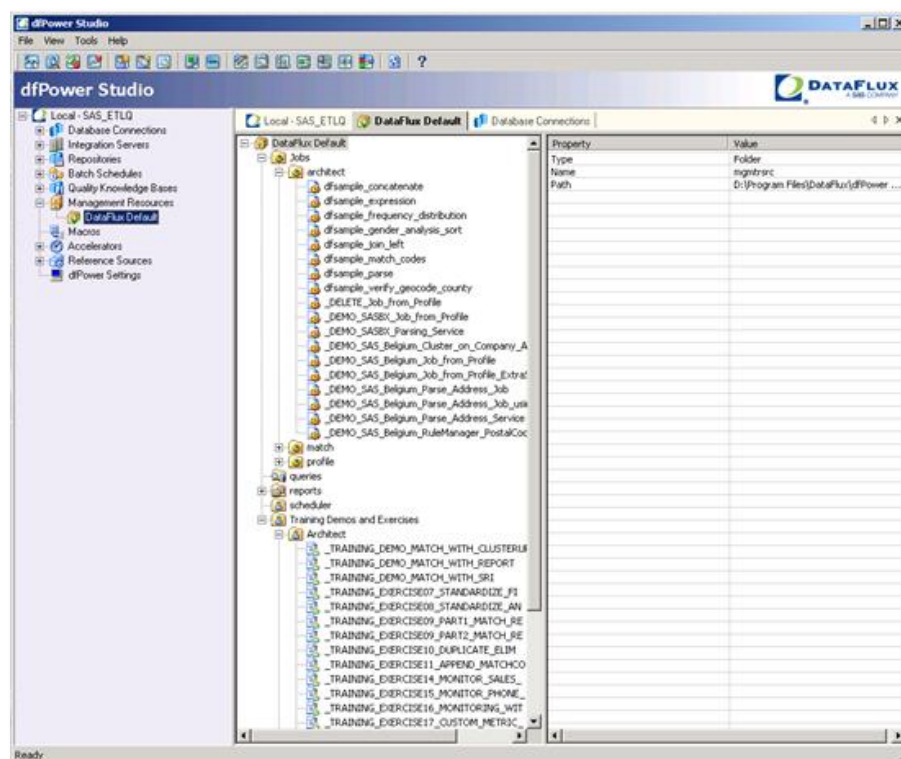


Figure 1. dfPower Studio Navigator

DATA ACCESS

All of the new features described in this section are available in both dfPower Studio and DataFlux Integration Server.

Starting with DataFlux release 8.0, the complete system is now Unicode-enabled. This means that all character sets are now supported when reading and writing files and databases and that these character sets can be displayed in the user interface. The latter requires that a Unicode-enabled font is used.

Direct Access to SAS data sets has been enhanced. You can now read and write SAS data sets directly without using a SAS server or SAS software. Also, Cross-Environment Data Access (CEDA) is supported when reading SAS data sets. This enables you, for example, to read a SAS data set using DataFlux software that runs on Microsoft Windows while the data set was created by SAS running on UNIX.

Starting with DataFlux release 8.2, data connections to SAS data sets are managed in a central location in dfPower Studio Navigator and can then be used in all DataFlux functionality, including dfPower Profile, dfPower Architect, and


```

Cobol_Copybook_with_Level_88.cpy - Notepad
File Edit Format View Help
***** Top of Data *****
* DataFlux Copybook Sample *
*****
01  OUT1-RECORD.
   05  OUT1-HEADER.
       10  OUT1-ID                PIC 9(5)  COMP-3.
       10  OUT1-OCCUR-LENGTH     PIC 9(5)  COMP-3.
   05  M-DMG-INFO.
       10  OUT1-TXT              PIC X(16).
       10  OUT2-TXT              PIC X(16).
       10  OUT3-TXT              PIC X(16).

01  INPUT-STATUS-FLAG           PIC X(01).
   88  INPUT-FILE-UNOPENED     VALUE 'A'.
   88  INPUT-FILE-UNREAD       VALUE 'B'.
   88  INPUT-FILE-STARTED      VALUE 'C'.
   88  INPUT-FILE-EOF          VALUE 'D'.

01  FILLER                      PIC X(25).
    
```

Figure 4. Example of COBOL Copybook

The screenshot shows the DataFlux COBOL Copybook Layout Editor interface. On the left, a tree view displays the hierarchy of the copybook: Copybook, OUT1-RECORD, OUT1-HEADER, OUT1-ID, OUT1-OCCUR-LENGTH, M-DMG-INFO, OUT1-TXT, OUT2-TXT, OUT3-TXT, INPUT-STATUS-FLAG, and FILLER. The 'INPUT-STATUS-FLAG' field is selected. On the right, a table displays the properties for this field.

Name	Value
Name	INPUT-STATUS-FLAG
Description	
Output Name	
Level Number	1
Occurs Maximum Value	0
Occurs Dependency	
Redefine	
Field Type	Field
Output Length	5
Picture	X(01)
Constant Value	
Usage	DISPLAY
Sign Leading	0
Sign Separate	0
INPUT-FILE-UNOPENED	A
INPUT-FILE-UNREAD	B
INPUT-FILE-STARTED	C
INPUT-FILE-EOF	D

Figure 5. DataFlux COBOL Copybook Layout Editor

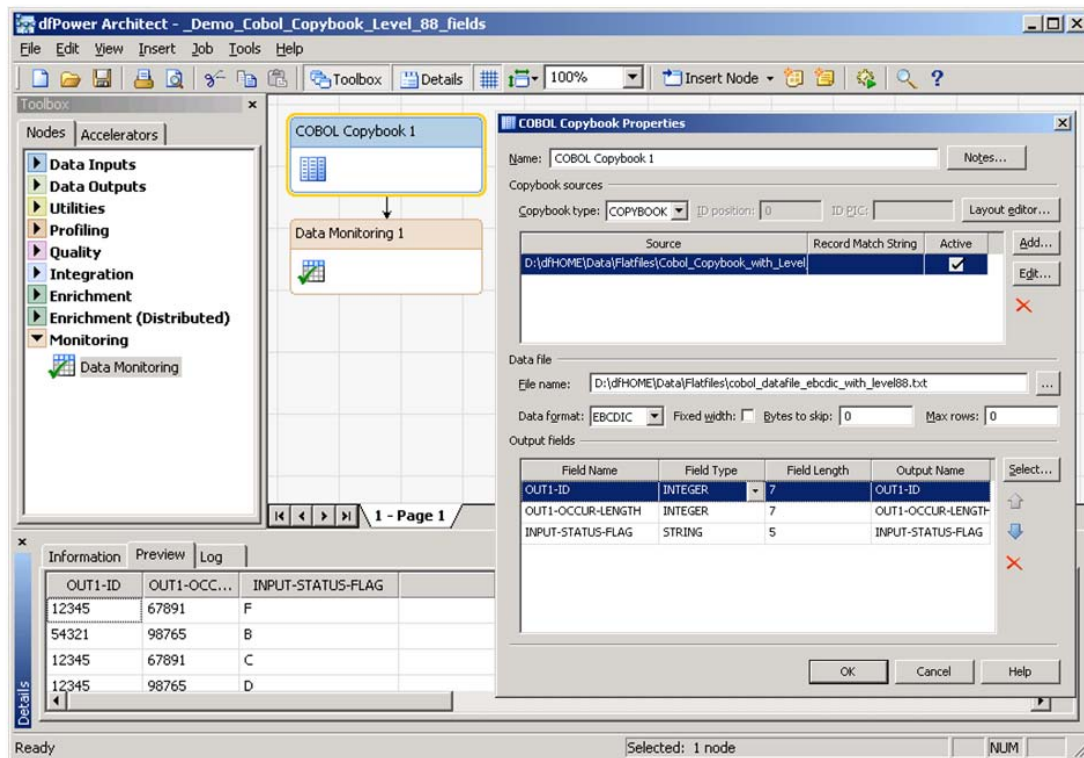


Figure 6. COBOL Copybook Node in dfPower Architect

New XML Input and XML Output nodes have been added in DataFlux release 8.2 to enhance support for XML. These nodes allow for reading and writing more complex XML structures and also support a number of industry-standard XML formats, such as X12, ATIS, HIPAA, EDIFACT, EDIGAS, NCPDP and TRADACOMS, EANCOM, IATA, HL7.

NEW FUNCTIONALITY IN dfPOWER ARCHITECT

dfPower Architect enables you to interactively design your data quality flows and to execute them locally or on a DataFlux Integration Server. As already mentioned in the section about general UI improvements, a number of job editing functionalities in dfPower Architect have been enhanced to increase productivity. These include:

- Copy/Paste support for node(s)
- Auto-Arrange nodes of nodes on a page
- Find node on page
- Ability to control where a note is inserted
- Ability to move multiple fields in the node output list
- Undo/Redo support (release 8.2 and above)

Also, a number of new nodes have been added. All the new nodes described in this section can be used in jobs that are executed locally using dfPower Architect or can be executed by DataFlux Integration Server as part of jobs and Web services.

New nodes that have been added include:

- Job Specific Data
- Email and FTP
- Execute External Program
- Repository Info
- Repository Detail

The Job Specific Data node enables you to embed data inside your job for testing purposes. For example, when developing an Embedded Job or a Web service, you would temporarily add a Job Specific Data Node to simulate the input values, while having access to all the debugging facilities to go through the processing step by step.

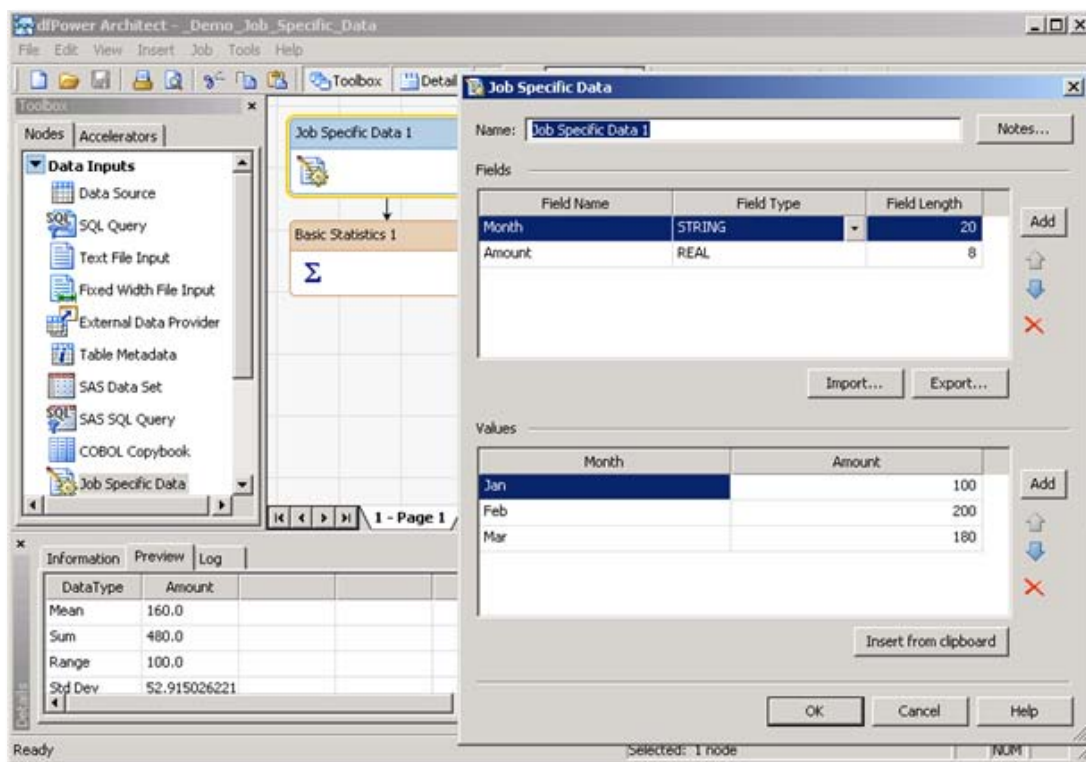


Figure 7. Using the Job Specific Data Node

The Email and FTP node enables you to send emails or files as part of your job flow. The Execute External Program node enables you to execute any third-party program that has a command-line interface and enables you to pass data from fields in your job flow to STDIN and to retrieve information from STDOUT that was generated by the third-party program. This allows for easy integration of third-party programs without having to store the results into a temporary file.

The Repository Info and Repository Detail node enable you to extract information from DataFlux repositories. This enables you to extract results that were generated by applying business rules using the dfPower Monitor functionality.

Furthermore, a number of existing nodes have been enhanced. For example, the Parsing node and the Standardization node now support a Result Code that indicates whether the operation was successful or not. More information can be found in the Changes & Enhancements section of the online Help.

PERFORMANCE

Threaded sorting is now supported in both dfPower Studio and DataFlux Integration Server. This enables the software to take advantage of multiple CPUs to speed up data processing. There are configuration options available to specify memory usage for each thread and to specify the maximum number of threads.

Threaded sorting enhances the performance of the following nodes:

Data Access capabilities have been enhanced by providing support for COBOL copybooks for use in data quality initiatives for legacy systems. Enhanced support for XML, including support for various industry standards, and enhanced support for SAS data sets, allows for easy integration of these data sources in your data quality initiatives.

Support for multi-threading enables you makes it possible to take better advantage of multi-CPU systems and to process the ever-growing amounts of data more quickly.

In summary, there are many new reasons to start using the latest version of DataFlux software for your data quality initiatives to help you make your data quality initiatives easier, faster, and more productive.

CONTACT INFORMATION

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

Wilbram Hazejager
DataFlux Corporation
Heidelberg, Germany
Work Phone: +49-6221-4150
E-mail: Wilbram.Hazejager@dataflux.com
Web: www.dataflux.com

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.