Paper 141-2011

Innovations in Data Management: Introduction to DataFlux® Data Management Platform

Wilbram Hazejager, DataFlux Corporation, Cary, NC

ABSTRACT

DataFlux®, a SAS® company, recently delivered the first installment of the DataFlux® Data Management Platform in order to provide complete access to enterprise data. This set of products, which will be included with many SAS® 9.3 product offerings, covers the gamut of data management needs including data quality, data integration, and master data management. In this presentation, we provide a view of the fundamental product capabilities delivered with the DataFlux Data Management Platform and show how a targeted user interface can make even the most challenging problems simple to solve.

INTRODUCTION

The DataFlux Data Management Platform is an integrated data management technology that enables organizations to seamlessly manage the quality of source information, as well as integrate information from complex, disparate data sources into one single, unified view. The platform offers a single, integrated technology suite that addresses every phase of the data quality and data integration life cycle, including data profiling, cleansing, consolidation, enrichment, and monitoring. The solution can be deployed in both batch and real-time environments, as well as directly within many operational systems and Extract, Transform, and Load (ETL) applications.

The platform consists of a number of components as displayed in the following figure.

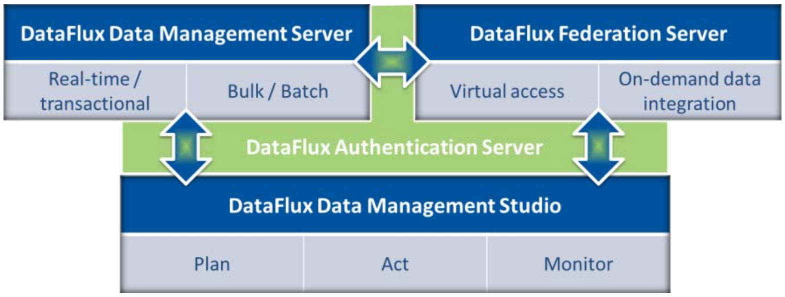

DataFlux Data Management Server
Real-time / transactional
Bulk / Batch
DataFlux Federation Server
Virtual access
On-demand data integration
DataFlux Authentication Server
DataFlux Data Management Studio
Plan
Act
Monitor

Figure 1: Components of DataFlux Data Management Platform

DataFlux® Data Management Studio provides a single interface for both business and IT users to plan, implement, and monitor the rules to manage data throughout the organization. This three-stage business methodology approach – Plan, Act, and Monitor – is part of the DataFlux methodology which supports the entire data integration life cycle through an integrated phased approach.

DataFlux® Data Management Server supports the ability to run batch jobs that were developed using DataFlux Data Management Studio. DataFlux Data Management Server is available in three editions: DataFlux® Standard Data Management Server, DataFlux® Enterprise Data Management Server, and DataFlux® Data Management Server for SAS.

DataFlux Standard Data Management Server enables any DataFlux Data Management Studio user to offload jobs to a more scalable server environment.

DataFlux Data Management Server for SAS (that will be available with SAS 9.3 and later) additionally enables SAS programs to invoke DataFlux Data Management Platform batch jobs and business services from a SAS program.

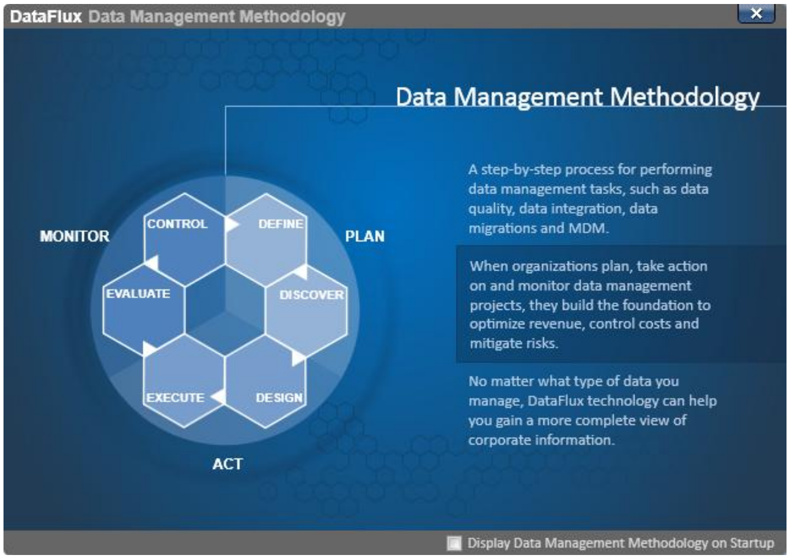
DataFlux Enterprise Data Management Server has added the ability to call DataFlux business services that were designed in the DataFlux Data Management Studio environment using a service-oriented architecture (SOA) from any third-party application.

DataFlux® Federation Server provides federated data access. Additionally, it provides SAS users with a data server for SAS data sets to be used in combination with DataFlux software.

DataFlux® Authentication Server is used by DataFlux servers to define and control access to these servers.

THE DATAFLUX METHODOLOGY

The DataFlux methodology supports the entire data integration life cycle through an integrated phased approach. These phases include data profiling, data quality, data integration, data enrichment, and data monitoring. The methodology may be implemented as an ongoing process to manage data as well as cleanse and improve data throughout the enterprise.



Data Management Methodology

A step-by-step process for performing data management tasks, such as data quality, data integration, data migrations and MDM.

When organizations plan, take action on and monitor data management projects, they build the foundation to optimize revenue, control costs and mitigate risks.

No matter what type of data you manage, DataFlux technology can help you gain a more complete view of corporate information.

□ Display Data Management Methodology on Startup

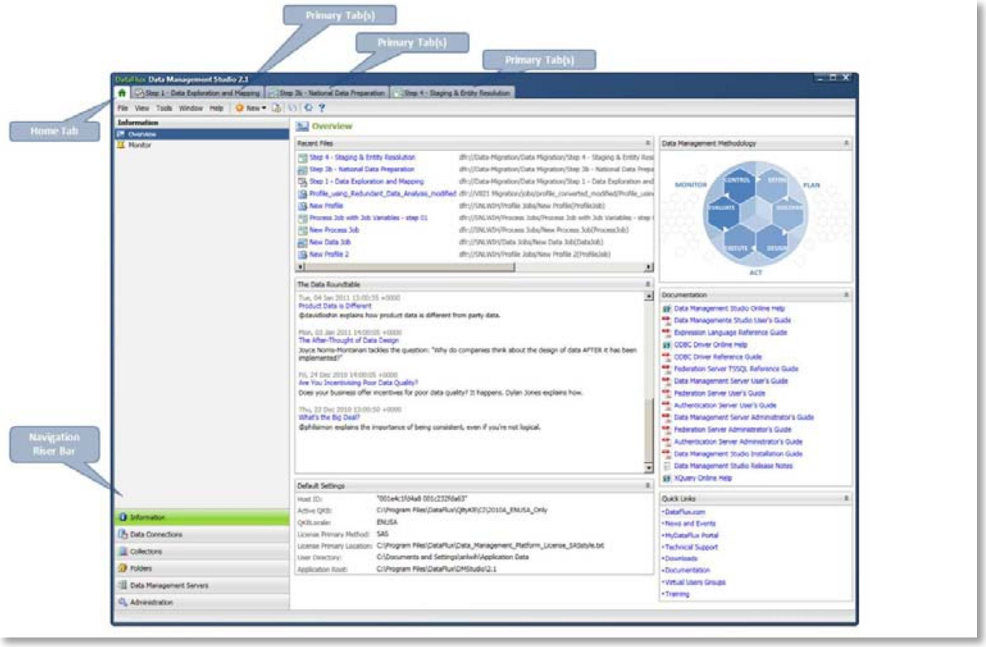
Figure 2: DataFlux Data Management Methodology

Additionally, this methodology fits into an overarching three-stage business methodology approach – Plan, Act, and Monitor. The first stage of this methodology, Plan, focuses solely on data discovery or assessment to accurately identify the consistency, accuracy, and validity of the source data. During this stage, data quality issues are identified and documented, and business rules are created to correct the data quality issues. The second stage, Act, supports the flexible correction of the identified data quality issues and if appropriate, the improvement of core business processes. The last stage, Monitor, supports ongoing monitoring and trending of source data to ensure information accuracy and to automatically detect and alert users if data violates defined business rules or corporate data quality standards.

DataFlux Data Management Studio contains an interactive wizard that shows and explains the various steps of the methodology. From the wizard, the user can jump directly to functionality inside Data Management Studio that can be used to support the specific step in the methodology. The wizard is optional and does not force the user to go through the various steps in a pre-defined order. The wizard should be considered a guideline and enables first time and occasional users to get started quickly.

DATAFLUX DATA MANAGEMENT STUDIO

The main user interface has been designed using the latest standards to make it more intuitive to use for first time and occasional users without limiting the power user. Navigation riser bars are used in many places. A tabbed display is used instead of opening separate windows.



Recent Files

Step 4 - Staging & Entity Resolution	dfr://Data-Migration/Data Migration/Step 4 - Staging & Entity Resol
Step 3b - National Data Preparation	dfr://Data-Migration/Data Migration/Step 3b - National Data Prepa
Step 1 - Data Exploration and Mapping	dfr://Data-Migration/Data Migration/Step 1 - Data Exploration and
Profile_using_Redundant_Data_Analysis_modified	dfr://V821 Migration/jobs/profile_converted_modified/Profile_usin
New Profile	dfr://SNLWDH/Profile Jobs/New Profile(ProfileJob)
Process Job with Job Variables - step 01	dfr://SNLWDH/Process Jobs/Process Job with Job Variables - step
New Process Job	dfr://SNLWDH/Process Jobs/New Process Job(ProcessJob)
New Data Job	dfr://SNLWDH/Data Jobs/New Data Job(DataJob)
New Profile 2	dfr://SNLWDH/Profile Jobs/New Profile 2(ProfileJob)

The Data Roundtable

Tue, 04 Jan 2011 13:00:35 +0000
Product Data is Different
@davidloshin explains how product data is different from party data.

Mon, 03 Jan 2011 14:00:05 +0000
The After-Thought of Data Design
Joyce Norris-Montanari tackles the question: "Why do companies think about the design of data AFTER it has been implemented?"

Fri, 24 Dec 2010 14:00:05 +0000
Are You Incentivising Poor Data Quality?
Does your business offer incentives for poor data quality? It happens. Dylan Jones explains how.

Thu, 23 Dec 2010 13:00:50 +0000
What's the Big Deal?
@philsimon explains the importance of being consistent, even if you're not logical.

Documentation

- Data Management Studio Online Help
- Data Managements Studio User's Guide
- Expression Language Reference Guide
- ODBC Driver Online Help
- ODBC Driver Reference Guide
- Federation Server TSSQL Reference Guide
- Data Management Server User's Guide
- Federation Server User's Guide
- Authentication Server User's Guide
- Data Management Server Administrator's Guide
- Federation Server Administrator's Guide
- Authentication Server Administrator's Guide
- Data Management Studio Installation Guide
- Data Management Studio Release Notes
- XQuery Online Help

Default Settings

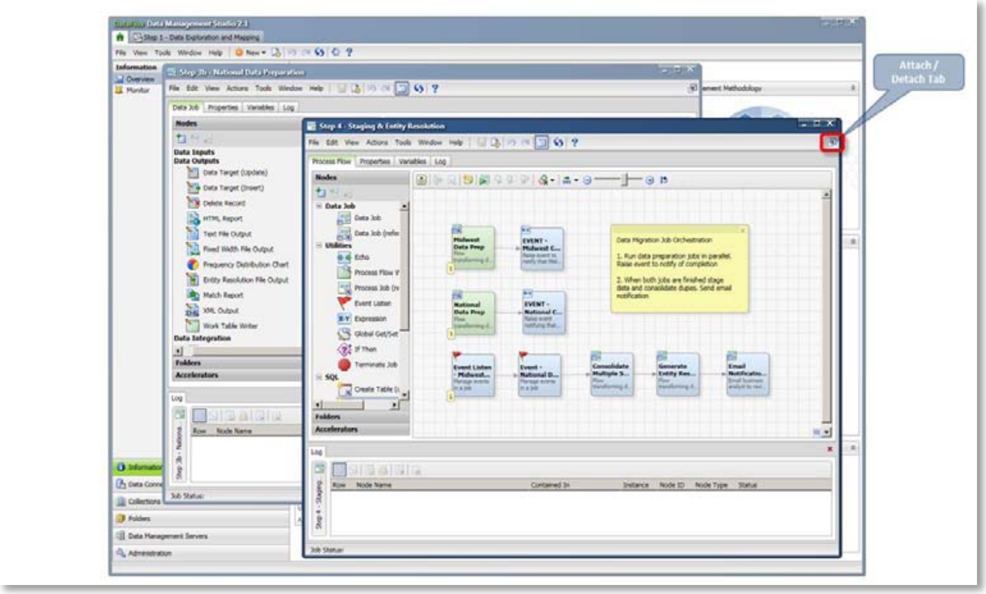
Host ID:	"001e4c1f04a8 001c232f6a63"
Active QKB:	C:\Program Files\DataFlux\QltyKB\CI\2010A_ENUSA_Only
QKB Locale:	ENUSA
License Primary Method:	SAS
License Primary Location:	C:\Program Files\DataFlux\Data_Management_Platform_License_SASstyle.txt
User Directory:	C:\Documents and Settings\ankwh\Application Data
Application Root:	C:\Program Files\DataFlux\DMStudio\2.1

Quick Links

- DataFlux.com
- News and Events
- MyDataFlux Portal
- Technical Support
- Downloads
- Documentation
- Virtual Users Groups
- Training

Figure 3: DataFlux Data Management Studio User Interface

The primary tabs can be detached to show the information in multiple windows side-by-side, which can be very efficient for a power user.



Data Migration Job Orchestration

1. Run data preparation jobs in parallel. Raise event to notify of completion

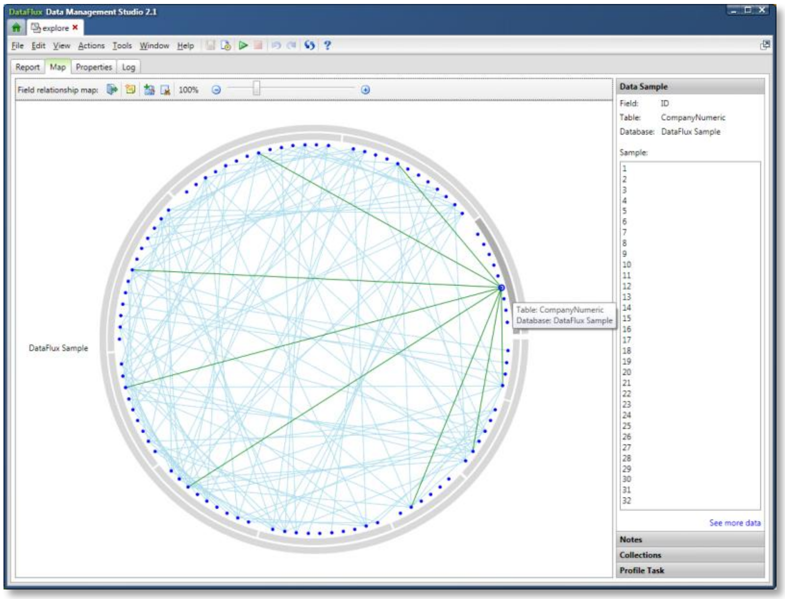
2. When both jobs are finished stage data and consolidate dupes. Send email notification

Figure 4: Detached Tabs in the DataFlux Data Management Studio User Interface

DATA EXPLORATION

DataFlux exploration profiles databases to discover the metadata in hundreds - or thousands - of data sources. This profile helps to streamline the process of starting mission-critical data improvement projects.

When starting any data quality or data integration project, metadata is an invaluable tool for establishing which data to include in the initiative. This information is then used to find fields with similar names (using fuzzy matching techniques). The software also looks at data samples from each field to identify the type of data. The results can be generated into a report that can be used to identify potential redundancies and relationships.



Field: ID
Table: CompanyNumeric
Database: DataFlux Sample
Sample:
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

Figure 5: DataFlux Exploration Map

In the exploration map, the outer ring represents data sources. The elements in the inner ring represent the tables in each data source, and the dots represent the field names in each table.

For example, metadata about customer names can be grouped into Collections. DataFlux Data Management Platform introduces the concept of collections to allow users to bucket fields for quicker viewing and reporting. Users can create and add to a collection while exploring or profiling the data as well as through the Collections Riser.

Part of the power of collections is evident when viewing profile reports. Instead of navigating through a large amount of data connections and tables that might make up one report, users can switch to the Collection view which displays the grouping of the fields in one easy to see place.

LINEAGE

One of the key pieces to know during the Control phase is to know the impact of a change. DataFlux Data Management Studio helps a user understand the lineage of a particular artifact.

For every object that DataFlux Data Management Platform manages, a user can see the lineage of the object. The detail pane on the left shows the artifacts that are consumed as well as what consumes the selected object. The following image gives an example of lineage.

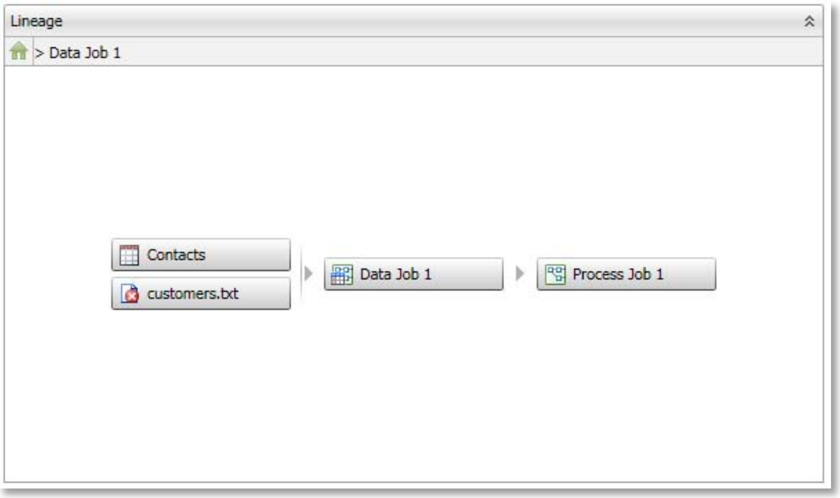


Figure 6: An Example of Data Lineage from DataFlux Data Management Platform

The lineage diagram on Data Job 1 shows that this job is using a data source called Contacts. It also uses a flat file called customers.txt. The red x indicates that this flat file currently does not exist. On the right side of Data Job 1, we can see that the data is used in Process Job 1. Click on any of the objects in this lineage diagram to show the lineage for the selected object. Double-clicking on an object opens the properties of that object, or in case of a job, the job flow diagram opens.

From one simple interface, users can quickly navigate the dependency tree to see how well business rules are being used, or if a change is needed, all the areas that need to be reviewed after the change is made.

PROCESS JOBS

Process jobs in DataFlux Data Management Platform allow users to implement advanced process logic. For example, process jobs contain support for event handling and parallelization.

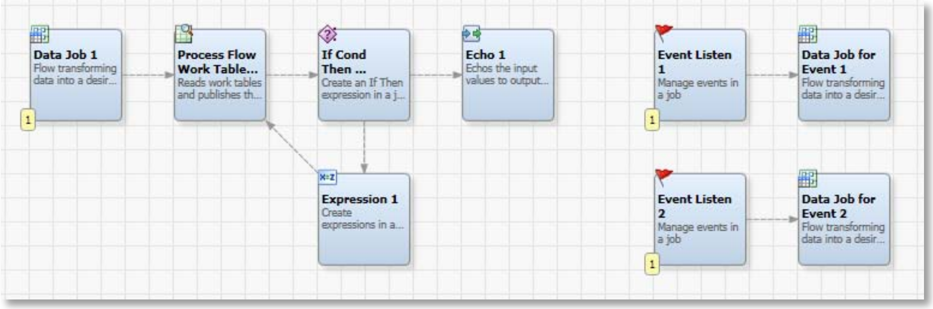
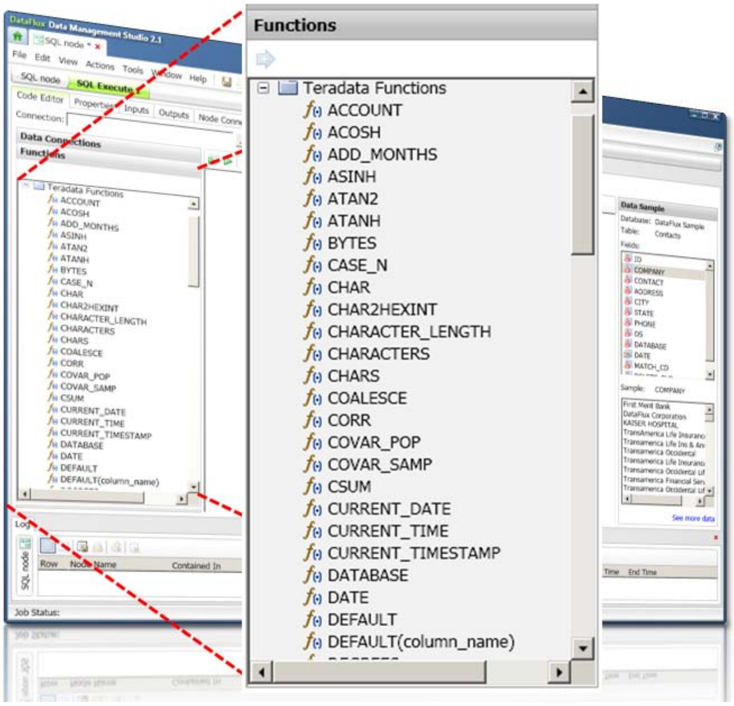

Data Job 1
Flow transforming data into a desir...
1
Process Flow Work Table...
Reads work tables and publishes th...
If Cond Then ...
Create an If Then expression in a j...
Echo 1
Echos the input values to output...
Event Listen 1
Manage events in a job
1
Data Job for Event 1
Flow transforming data into a desir...
Expression 1
Create expressions in a...
Event Listen 2
Manage events in a job
1
Data Job for Event 2
Flow transforming data into a desir...

Figure 7: An Example of a Process Job with Event Handling

In cases where several steps need to take place and complete before the next step can start, the event handling capability in DataFlux Data Management Platform makes this easier to accomplish. This type of task was originally done through complex scripts or writing to and reading from files to make sure that all pre-tasks had completed before processing the next step. In a process job, the complexity is greatly reduced, and users can easily manage what needs to happen and when it needs to happen.

In a process job, users can set up the execution of the same node or collection of nodes to run at the same time. Some advantages of this are reducing the time to process larger amounts of data as well as being able to run multiple prerequisite steps all at the same time. The result is improved job performance. Combined with event handling, parallelization is a powerful feature that provides unique processing control at the user's fingertips.

A powerful SQL node allows users to build the most complex SQL queries (that can include database specific functionality) and push the query down to the database. This processing node provides a user interface that enables the user to build their query interactively and also lists the database specific functions.



Functions

Teradata Functions
- ACCOUNT
- ACOSH
- ADD_MONTHS
- ASINH
- ATAN2
- ATANH
- BYTES
- CASE_N
- CHAR
- CHAR2HEXINT
- CHARACTER_LENGTH
- CHARACTERS
- CHARS
- COALESCE
- CORR
- COVAR_POP
- COVAR_SAMP
- CSUM
- CURRENT_DATE
- CURRENT_TIME
- CURRENT_TIMESTAMP
- DATABASE
- DATE
- DEFAULT
- DEFAULT(column_name)

DataFlux Data Management Studio 2.1
SQL node
File Edit View Actions Tools Window Help
SQL Execute
Code Editor Properties Inputs Outputs Node Connections
Data Connections
Data Sample
Log
Job Status:

Figure 8: ELT Support

This ELT support allows the database to optimize the query and run it inside the database, which removes the need for data movement and further improves job performance.

ENTITY RESOLUTION

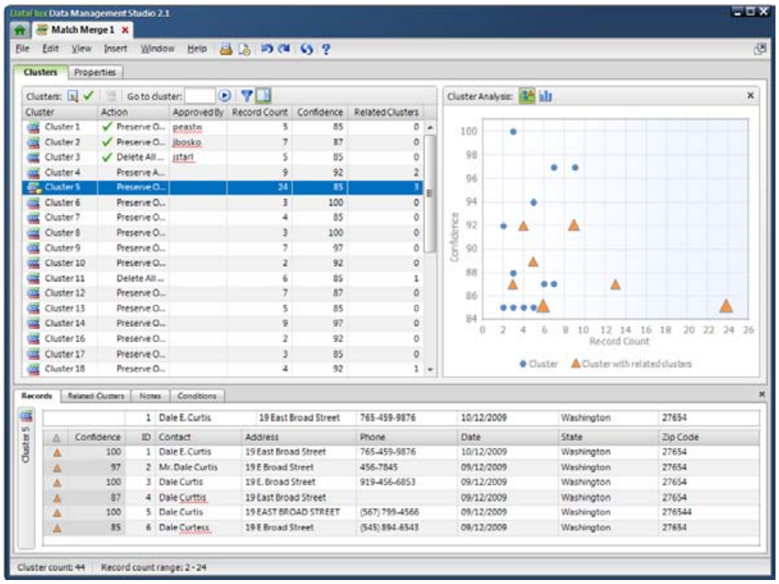
Entity resolution is the process of merging multiple files (or duplicate records within a single file) in such a way that records referring to the same physical object are treated as a single record. Records are matched based on the information that they have in common. The records that you can merge appear to be different but can actually refer to the same person or thing.

The DataFlux match engine has been designed to enable both the identification of duplicate records within a single data source, as well as across multiple sources. The rules-based matching engine uses a combination of parsing rules, standardization rules, phonetic matching, and token-based weighting to strip the ambiguity out of source information.

After matching has been performed and clusters are created, record consolidation (also called duplicate elimination) merges multiple records into a single "best record.” Through user-defined “record-level” and “field-level” rules, the engine is able to pick and choose information from multiple records in order to compile a single version of the entity.

For example, a record-level rule may call for the preservation of a record with the most recent edit or create date. However, this record may not include accurate address information. If the address exists in another record, field level rules can be used to extract the address from the secondary record and then replace the address in the primary record with this trusted content.

Besides a fully automated process, an interactive Match Reviewer is provided to review the results and if necessary, modify the match, clustering, and best-record-selection rules.



Cluster	Action	Approved By	Record Count	Confidence	Related Clusters
Cluster 1	✓ Preserve O...	peastw	5	85	0
Cluster 2	✓ Preserve O...	jbosko	7	87	0
Cluster 3	✓ Delete All ...	jstart	5	85	0
Cluster 4	Preserve A...		9	92	2
Cluster 5	Preserve O...		24	85	3
Cluster 6	Preserve O...		3	100	0
Cluster 7	Preserve O...		4	85	0
Cluster 8	Preserve O...		3	100	0
Cluster 9	Preserve O...		7	97	0
Cluster 10	Preserve O...		2	92	0
Cluster 11	Delete All ...		6	85	1
Cluster 12	Preserve O...		7	87	0
Cluster 13	Preserve O...		5	85	0
Cluster 14	Preserve O...		9	97	0
Cluster 16	Preserve O...		2	92	0
Cluster 17	Preserve O...		3	85	0
Cluster 18	Preserve O...		4	92	1

		1	Dale E. Curtis	19 East Broad Street	765-459-9876	10/12/2009	Washington	27654
▲	Confidence	ID	Contact	Address	Phone	Date	State	Zip Code
▲	100	1	Dale E. Curtis	19 East Broad Street	765-459-9876	10/12/2009	Washington	27654
▲	97	2	Mr. Dale Curtis	19 E Broad Street	456-7845	09/12/2009	Washington	27654
▲	100	3	Dale Curtis	19 E. Broad Street	919-456-6853	09/12/2009	Washington	27654
▲	87	4	Dale Curttis	19 East Broad Street		09/12/2009	Washington	27654
▲	100	5	Dale Curtis	19 EAST BROAD STREET	(567) 799-4566	09/12/2009	Washington	276544
▲	85	6	Dale Curtess	19 E Broad Street	(545) 894-6543	09/12/2009	Washington	27654

Cluster count: 44 Record count range: 2 - 24

Figure 9: Match Reviewer

DATAFLUX FEDERATION SERVER

The DataFlux Federation Server is a key component to the data security strategy in Data Management Platform. The key functionality of the Federation Server is allowing IT administrators to setup data sources in one central place that can be used by DataFlux software. The IT administrators can apply security at the server layer that enhances what is already in place at the database level. Administrators can control who can see what data sources and what actions they can do. Even if a data source provides the capability to update data, a user can be denied this action from within the Federation Server.

The Federation Server also allows customers to use native data drivers to access data. With Data Management Studio and Data Management Server, the only options are the ODBC drivers. The Federation Server opens this up to include native drivers that the server supports.

Additionally, the Federation Server can act as a data server for SAS data sets and provides a security layer for accessing these SAS data sets.

HOW TO MIGRATE TO THE NEW PLATFORM?

The new DataFlux Data Management Platform provides wizards that enable you to convert your DataFlux® dfPower® Studio repositories, DataFlux® dfPower® Profile jobs, and dfPower Architect jobs to equivalent Data Management Platform objects.

A recommended approach is first to migrate a dfPower Studio repository and any related Management Resources. Then, migrate any additional jobs later. After a repository is migrated, some queries, business rules, custom metrics, and related items will be ready to use in Data Management Studio. Use the wizards to make your dfPower Architect and dfPower Profile jobs available for use Data Management Studio. If jobs take advantage of macros or use a Merge File Output node, then the job needs to be edited after conversion due to improvements around macro resolution and repository structures in Data Management Platform. The *DataFlux Migration Guide* with a detailed description of the process is available on the MyDataFlux Portal (www.dataflux.com).

SAS customers will be able to acquire the new DataFlux Data Management Platform as part of SAS 9.3.

CONCLUSION

This paper has highlighted some of the capabilities in the new DataFlux Data Management Platform. This innovative data management technology enables business agility and IT efficiency to help organizations manage critical data in the areas of data quality, data integration, and master data management (MDM).

By combining these capabilities into a unified platform, DataFlux helps companies deliver reliable, trusted data across the enterprise.

More information about DataFlux Data Management Platform can be found at http://www.dataflux.com/Products/Products.aspx.

CONTACT INFORMATION

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

Wilbram Hazejager
DataFlux Corporation
Cary, NC

E-mail: Wilbram.Hazejager@dataflux.com