# DataFlux® Data Management Studio Installation and Configuration Guide

sas®

DATAFLUX®
data management

This page is intentionally blank

# DataFlux® Data Management Studio

## Installation and Configuration Guide

Version 2.4

December 18, 2012

This page is intentionally blank

# Contact DataFlux

**DataFlux Corporate Headquarters**
Toll Free: (877) 846-3589
Tel: (919) 447-3000
Fax: (919) 447-3100
940 NW Cary Parkway, Suite 201
Cary, NC 27513
USA

**DataFlux West**
Tel: (818) 906-7638
Fax: (818) 907-6012
15300 Ventura Boulevard, Suite 523
Sherman Oaks, CA 91403
USA

## Technical Support

Phone: 919-677-8008
Email: techsupport@sas.com
Web: http://support.sas.com/techsup/contact/

## Documentation Support

Email: yourturn@sas.com

## Legal Notices

# Table of Contents

# Introduction



## Accessibility

DataFlux Data Management Platform (Data Management Platform) includes features that improve usability of the product for users with disabilities. These features are related to accessibility standards for electronic information technology that were adopted by the United States (U.S.) Government under Section 508 of the U.S. Rehabilitation Act of 1973, as amended.

If you have questions or concerns about the accessibility of DataFlux products, send an e-mail to techsupport@dataflux.com.

## Conventions Used in this Document

This document uses several conventions for special terms and actions.

### Typographical Conventions

The following typographical conventions are used in this document:

| Typeface | Description |
|---|---|
| **Bold** | Text in bold signifies a button or action |
| *italic* | Identifies document and topic titles |
| `monospace` | Typeface used to indicate filenames, directory paths, and examples of code |

### Syntax Conventions

The following syntax conventions are used in this document:

| Syntax | Description |
|---|---|
| [] | Brackets [] are used to indicate variable text, such as version numbers |
| # | The pound # sign at the beginning of example code indicates a comment that is not part of the code |
| > | The greater than symbol is used to show a browse path, for example **Start** > **Programs** > **DataFlux Data Management Studio** 2.2 > **Documentation**. |

# Reference Publications

This document might reference other DataFlux® publications including:

*DataFlux Authentication Server Administrator's Guide*

*DataFlux Authentication Server User's Guide*

*DataFlux Data Management Server Administrator's Guide*

*DataFlux Data Management Server User's Guide*

*DataFlux Data Management Studio User's Guide*

*DataFlux Expression Language Reference Guide*

*DataFlux Federation Server Administrator's Guide*

*DataFlux Federation Server User's Guide*

*DataFlux Migration Guide*

DataFlux Quality Knowledge Base Online Help

# Installing Data Management Studio



## Installing Data Management Studio

DataFlux Data Management Studio is available through SAS delivery channels. See your SAS Software Order Email (SOE) for information about installing this product.

The default installation path under Windows is: SASHome\<product_instance_name>

The default installation path under UNIX is: SASHome/<product_instance_name>

If you want to configure Data Management Studio after installation, see Configuring Data Management Studio. If you want to install add-on products, see Installing Add-On Products.

## System Requirements

You can review system requirements for SAS 9.3 products at the following location:

http://support.sas.com/documentation/installcenter/93/

A search window for SAS 9.3 Install Center Documentation appears. Search for your product name. A results page will appear with links to the system requirements for your software.

## Supported Databases for Data Storage

DataFlux Data Management Studio can use DataFlux ODBC 7.0 drivers to access data in the following databases:

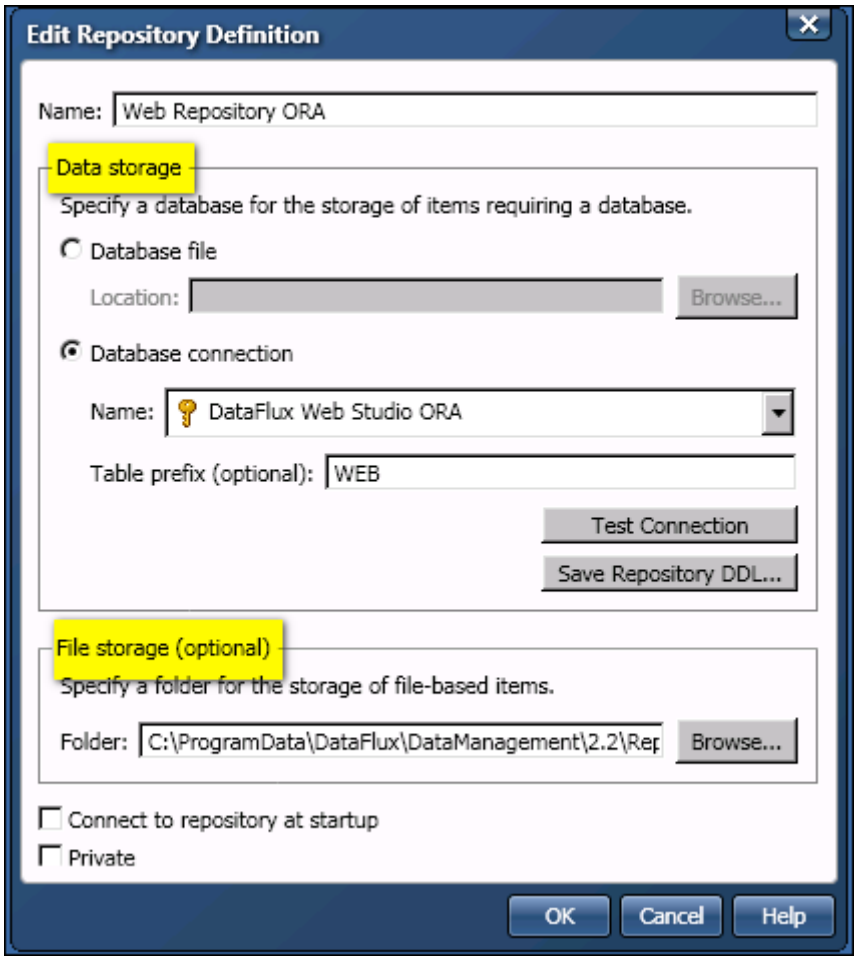
| Driver | Database Version |
|---|---|
| Btrieve® | Btrieve 6.15 |
| Btrieve | Pervasive.SQL™ 7.0, 8.5, and 2000 |
| dBASE | Clipper™ |
| dBASE™ | dBASE IV and V |
| dBASE | FoxPro 2.5, 2.6, and 3.0 |
| dBASE | FoxPro 6.0 (with 3.0 functionality only) |
| dBASE | FoxPro 3.0 Database Containers (DBC) |
| DB2® Wire Protocol | IBM® DB2 v9.1, v9.5, and v9.7 for Linux®, UNIX, and |

| Driver | Database Version |
|---|---|
| | Windows® |
| DB2 Wire Protocol | IBM DB2 Universal Database™ (UDB) v7.x and v8.x for Linux, UNIX, and Windows |
| DB2 Wire Protocol | IBM DB2 v9.1 and v10 for z/OS |
| DB2 Wire Protocol | IBM DB2 UDB v7.x and v8.1 for z/OS |
| DB2 Wire Protocol | IBM DB2 UDB V5R1, V5R2, V5R3, V5R4, V6R1, and V7R1 for iSeries® |
| Greenplum Wire Protocol | Greenplum Database 3.1, 3.2, 3.3, 4, and 4.1 |
| Informix© (client) | Informix Dynamic Server 9.2, 9.3, 9.4, 10.0, 11.0, 11.5, and 11.7 |
| Informix Wire Protocol | Informix Dynamic Server 9.2, 9.3, 9.4, 10.0, 11.0, 11.5, and 11.7 |
| MySQL™ Wire Protocol | MySQL 5.0x, 5.1, and 5.5 |
| Oracle® (client) | Oracle 8.0.5 and higher (32-bit only) |
| Oracle (client) | Oracle 8i R2 and R3, (8.1.6 and 8.1.7) |
| Oracle (client) | Oracle 9i R1 and R2 (9.0.1 and 9.2) |
| Oracle (client) | Oracle 10g R1 and R2 (10.1 and 10.2) |
| Oracle (client) | Oracle 11g R1 and R2 (11.1 and 11.2) |
| Oracle Wire Protocol | Oracle 8i R2, R3 (8.1.6 and 8.1.7) |
| Oracle Wire Protocol | Oracle 9i R1 and R2 (9.0.1 and 9.2) |
| Oracle Wire Protocol | Oracle 10g R1 and R2 (10.1 and 10.2) |
| Oracle Wire Protocol | Oracle 11g R2 (11.2) |
| PostgreSQL® Wire Protocol | PostgreSQL 8.2, 8.3, 8.4, 9.0, and 9.1 |
| Progress OpenEdge Wire Protocol | Progress OpenEdge 10.1x, 10.2x, and 11.0 |
| Salesforce | Salesforce (API Version 23) |
| SQL Server Legacy Wire Protocol | Microsoft® SQL Server® 7.0 |
| SQL Server Legacy Wire Protocol | Microsoft SQL Server 2000 |
| SQL Server Legacy Wire Protocol | Microsoft SQL Server 2005 |
| SQL Server Legacy Wire Protocol | Microsoft SQL Server 2008 R1, R2 |
| SQL Server Wire Protocol | Microsoft SQL Server 2000 |
| SQL Server Wire Protocol | Microsoft SQL Server 2005 |
| SQL Server Wire Protocol | Microsoft SQL Server 2008 R1, R2 |
| Sybase® Wire Protocol | Sybase Adaptive Server® 11.5 and 11.9 |
| Sybase Wire Protocol | Sybase Adaptive Server Enterprise® 12.0, 12.5x, 15.0, 15.5, and 15.7 |
| Sybase IQ | Sybase IQ 15.0, 15.1, 15.2, and 15.3 |
| Teradata® | Teradata 12.0, 13.0, and 13.1 |
| Teradata | Teradata V2R6.0, V2R6.1, and V2R6.2 |
| Text | Text Files |
| XML | XML Documents (tabular and hierarchical formatted) |

# Supported Databases for Repository Storage

## Overview

To add or update a repository, you will access the Repository Definition dialog. This dialog is used to create a repository definition file (.RCF file). There are two main sections of the Repository Definition dialog, as shown in the next figure:




**Data storage** - specifies a database for the storage of data explorations, profiles, and all objects in the Business Rule Manager (rules, tasks, custom metrics, etc.). Supported databases include SQLite and other database formats. For a list of supported formats, see Database Storage for Repositories.

**File storage** - specifies a separate storage location for objects that are stored as files, such as data jobs, process jobs, queries, *.sas files (SAS code files), and Entity Resolution Output files (*.sri files). For a list of supported formats, see File Storage for Repositories.

## Database Storage for Repositories

A Data Management Studio repository supports two kinds of storage: database storage and file storage. The following databases can be used for the database storage section of a repository:

| Database | ODBC Driver |
|---|---|
| DB2® V8.x, V9.1, V9.5, and V9.7 for Linux®, UNIX, and Windows | X |
| DB2® Universal Database (UDB) v7.x for Linux®, UNIX, and Windows | X |
| Informix Server 10.0, 11.0, and 11.5 | X |
| Microsoft® SQL Server® 2008 R1, R2 | X |
| Microsoft SQL Server 2005 | X |
| Microsoft SQL Server 2000 Enterprise Edition (64-bit) | X |
| Microsoft SQL Server 2000 Desktop Engine (MSDE 2000) | X |
| Microsoft SQL Server 2000 Service Packs 1, 2, 3, 3a and 4 | X |
| Oracle® 11g R1, R2 (11.1, 11.2) | X |
| Oracle 10g R1, R2(10.1, 10.2) | X |
| Oracle 9i R1, R2 (9.0.1, 9.2) | X |
| Sybase® Adaptive Server® 12.0, 12.5x, 15, and 15.5 | X |
| Teradata® 12.0 and 13.0 | X |

**Note:** Due to locking issues, file-based repositories and Microsoft® Access® based repositories are not recommended for production systems.

## File Storage for Repositories

The file storage location can be any physical path that is accessible to Data Management Studio, which is a Windows application. Any version of Windows that supports Data Management Studio can be used for file storage. For more information, see System Requirements.

It is possible to specify a UNIX location for file storage if this location has been made accessible to Data Management Studio as a network drive, etc. The following versions of UNIX are supported:

- AIX®: Version 5.3 and 6.1 on POWER® architectures
- HP-UX® PA-RISC: HP-UX 11iv2 (11.23), 11iv3 (11.31)
- HP-UX Itanium®: HP-UX 11iv2 (11.23), 11iv3 (11.31)
- Linux® for x86 (x86-32)
- Linux on x64
- Solaris™ on SPARC: Version 8, 9, 10
- Solaris on x64: Version 10

# Configuring Data Management Studio



## Configuration Files

When Data Management Studio starts, it will determine which configuration options are in effect by reading a series of configuration files, by looking in the environment, and by reading the command line. If there are two settings of the same name that exist in different configuration settings, the order in which the settings are read in determines which value is used. The last value read is used as the configuration setting.

Data Management Studio reads configuration settings in this order:

1. The app.cfg file in the etc folder where Studio is installed.
2. The app.cfg file in a user folder, such as drive:\Documents and Settings\USERNAME\Application Data\DataFlux\DataManagement\VERSION.
3. The application-specific configuration files in the etc folder, such as ui.cfg or dmserver.cfg.
4. The application-specific configuration files in a user folder.
5. The macros folder in the etc folder. The default path to the macros folder can be overridden with BASE/MACROS_PATH setting in the above configuration files.
6. The macros folder in a user folder.
7. The environment variables.
8. The command-line options if applicable.

# DataFlux Folder Permissions

The following table outlines the recommended operating system permissions for users of Data Management Studio.

| Directories | Users | Default Permissions |
|---|---|---|
| *DataFlux-home* | Administrators, Installer | Full control |
| *DataFlux-home*\Data Management Studio | Process user | Read and Execute, List Folder Contents |
| *DataFlux repositories (which can exist anywhere)* | Process user | Read and Execute, List Folder Contents |

Users need to be able to access the content and make changes to the DataFluxhome/etc location. Also, they should have access to their own user location located in %APPDATA%/DataFlux/.

# Configuration Options

The main configuration options for Data Management Studio are as follows:

| Option | Purpose | Source | Notes |
|---|---|---|---|
| **Base/General Application** | | | |
| ODBC_INI | Overrides the location of the odbc.ini file | | UNIX only. |
| BASE/AUTH_SERVER_LOC | Location of DataFlux Authentication Server | Optional | If specified, contains the IOM URI to an Authentication Server. The basic format of the Authentication Server IOM URI is `iom://<host>:<port>`. Where <host> is the name of the computer where the Authentication Server is executing and <port> is the port in which the Authentication Server may be contacted. The port should be specified as 21030 unless the default port for the Authentication Server has been changed. For more information, see the topic "Run a Job with Domain-Enabled ODBC Connections" in the Data Jobs chapter of the *DataFlux Data Management Studio User's Guide*. |
| BASE/LIBRARY_PATH | Path for Java jar dependencies | Optional | Determined by startup code (DFEXEC_HOME/lib). |
| BASE/PLUGIN_PATH | Path used by all | Optional | Determined by startup code. |

| Option | Purpose | Source | Notes |
|---|---|---|---|
| | subsystems to find plugins | | |
| BASE/EXE_PATH | Path containing executables | Optional | Calculated. |
| BASE/PRIMARY_LICENSE | Primary licensing method | Required by base | Must be set in the configuration file as DATAFLUX or SAS |
| BASE/PRIMARY_LICENSE_LOC | Location of the primary license file or server | Required by base | Must be set in the configuration file. |
| BASE/MACROS_PATH | Path for system macros.cfg file | Optional | If not specified, this file is located in the **etc** subfolder of the installation folder. |
| BASE/SECONDARY_LICENSE | Secondary licensing method | Required by base | Must be set in the configuration file as DATAFLUX or SAS. |
| BASE/SECONDARY_LICENSE_LOC | Location of the secondary license file or server | Required by base | Must be set in the configuration file. |
| BASE/LOGCONFIG_PATH | Full path to the log configuration file | Optional | Must be set in the configuration file or it defaults to logging.xml in the **etc** directory. |
| BASE/MESSAGE_PATH | Path to the message directory | Optional | Determined by startup code. |
| BASE/MESSAGE_LOCALE | Error message locale | Optional | If not specified, it is determined from the system locale. |
| BASE/MESSAGE_LEVEL | Error level of messages | Optional | 0 (or not specified) - normal messages; 1 - includes source file and line number in messages. |
| BASE/USER_PATH | Path for user configuration files | Optional | Automatically determined by dfcurver. |
| BASE/REPOS_DDL_LINE_PREFIX | Format the output of the DDL file that is generated for a repository from the Repository Definition dialog | Must be set in the configuration file. | For this macro and BASE/REPOS_DDL_LINE_SUFFIX only, specifying ^p as a value will cause a line break. |
| BASE/REPOS_DDL_LINE_SUFFIX | Format the output of the DDL file that is generated for a repository from the Repository Definition dialog. | Must be set in the configuration file. | For this macro and BASE/REPOS_DDL_LINE_PREFIX only, specifying ^p as a value will cause a line break. |
| BASE/REPOS_SYS_PATH | System path for repository configuration files | Optional | Automatically determined. |

| Option | Purpose | Source | Notes |
|---|---|---|---|
| BASE/REPOS_USER_PATH | User directory for repository configuration files | Optional | Automatically determined by dfcurver. |
| BASE/REPOS_FILE_ROOT | Overrides the root of the repository for URI lookups | Optional | If specified this is used as the root for repository when resolving the URI. In other words, the path in the URI will be concatenated to this path to give the actual filename of a URI. |
| BASE/REPOS_EVENT_WAIT_QUERYMS | Repository event that processes the wait time between processing queries. | Optional | Specifies how frequently in milliseconds to query the repository for changes in the event table. This might need to be changed due to slow servers or IT issues. This is an overriding value and the default is used if no value is set by the user. A setting of -1 disables events from client. |
| BASE/REPOS_EVENT_CLEAN_TIMEMIN | Repository event processor that removes all events older than X minutes before startup. | Optional | |
| BASE/TEMP | Temporary directory | Optional | If not specified, it inherits the value of the TEMP environment variable. |
| BASE/DATE_FORMAT | Specific date formats | Optional | If specified, it is iso8601. |
| BASE/TIME_BASE | Whether to use GMT time | Optional | If this is set to GMT (not the default), the current date returns in GMT. This will affect anything that uses the current date timestamp. |
| BASE/SORTBYTES | Specifies the bytes to use when sorting | Optional | |
| BASE/SORTMERGES | Enables merge during sort | Optional | |
| BASE/SORTTEMP | Specifies the temporary path for sorts | Optional | |
| BASE/SORTTHREADS | Specifies the number of sort threads | Optional | |
| BASE/EMAILCMD | Specifies the command used to send email | Required | Can include %T and %B where %T is replaced with the recipient and %B is a file containing the body of the message; also used by monitor event as well as architect nodes. |
| BASE/FTPGETCMD | Specifies the command used | Required | Should default in the install, as follows: |

| Option | Purpose | Source | Notes |
|---|---|---|---|
| | for Ftp Get Functionality | | • %U: Replace with username<br>• %P: Replace with password<br>• %S: Replace with server<br>• %T: Replace with local directory<br>• %F: Replace with Files to download, multiple separated by spaces<br>• %L: Replace with the log file to pipe the output. |
| BASE/FTPPUTCMD | Specifies the command used for Ftp Put Functionality | Required | BASE/FTPPUTCMD |
| BASE/LOGEXCEPTIONS | Exception logging | Optional | Exception logging defaults to off. Set this option to either 1 or a combination of letters. A setting of 1 lists the modules loaded at the time the exception occurred, some information on those modules, and the call stack that caused the error. A setting with letters can include: m = do not show module info, V=turn verbose on, U=install the Unhandled Exception Filter, C=install the continue Exception Filter, f=do not install the first chance exception filter. This must be set before starting the application of interest, as this setting is read only at startup. |
| BASE/TEXTMINE_LITI_LANG_LOCATION | Doc extraction node option | Optional | This is the install location of Teragram liti files. This allows them to be in Teragram provided languages instead of in the DataFlux install. |
| BASE/DMSTUDIO | Studio indicator | Optional | If this option is set to true (value of 1), it indicates that it is using the dmstudio process (not processes started by dmstudio, like dfwfproc for instance). The user should not adjust this or override the value. |
| BASE/APP_VER | Application version number | Optional | Defaults to 2.3. |
| BASE/UPDATE_LEVEL | Application update level | Optional | Defaults to 0. Could be used as a minor revision number. |
| **Data Access Component Logging** | | | |
| DAC/DISABLESYSCATENUM | Enumeration of syscat dsn's | Optional | When set to "yes", 1, or "true," this setting will disable the listing the |

| Option | Purpose | Source | Notes |
|---|---|---|---|
| | | | SYSCAT type DSN's into DSN's that are on that server. |
| DAC/DFTKLOGFILE | DFTK logging | Optional | Filename. |
| DAC/TKTSLOGFILE | TKTS logging | Optional | Filename. |
| DAC/DFTKDISABLECEDA | Disables CEDA support | Optional | "Yes" turns it on. |
| DAC/SAVEDCONNSYSTEM | Location of system-saved connections | Optional | Defaults to DFEXEC_HOME/etc/dsn. |
| DAC/SAVEDCONNUSER | Location of user-saved connections | Optional | Defaults to the user settings folder, the folder where all of the application-specific settings supplied by a user are stored, such as the following path under Windows 7: C:\Users\[username]\AppData\Roaming\DataFlux\dac\9.x |
| DAC/DSN | DSN directory for TKTS dsns | Optional | Path that defaults to DFEXEC_HOME/etc/dftkdsn. |
| DAC/DFTK_PROCESS | Run DFTK out of process | Optional | "Yes" turns it on; off by default. |
| DAC/DFTK_PROCESS_TKPATH | TKTS path for DFTK out of process | Optional | Path that defaults to a core/sasext dir off the executable dir. |
| **Address Update (NCOA) (in dfncoa_appcfg.h)** | | | |
| NCOA/DVDPATH | Path to the unpacked/unZIPped NCOA data | Required | Resides in macros/ncoa.cfg. |
| NCOA/QKBPATH | Path to the QKB used for Address Update name parsing | Required | Resides in macros/ncoa.cfg. |
| NCOA/USPSPATH | Path to the USPS CASS/DPV/etc data | Required | Resides in macros/ncoa.cfg. |
| NCOA/QKBPARSEDEFN | Name of the Addr. Updt. QKB parse definition to use | Optional | Default is "Name (Address Update)". Resides in macros/ncoa.cfg. |
| NCOA/REPOSDSN | Specifies DSN used to connect to the Address Update repository | Required | Is overridden by NCOA/REPOSCONNECTION. One or the other is required. This is typically set by the Address Update Admin utility. Resides in app.cfg. |
| NCOA/REPOSCONNECTION | Specifies the connection string used to connect to the Address Update repository | Required | Overrides NCOA/REPOSDSN. One or the other is required. This is typically set by the Address Update Admin utility. Resides in app.cfg. |

| Option | Purpose | Source | Notes |
|---|---|---|---|
| NCOA/REPOSPREFIX | Table prefix used on the Address Update tables. | Required | This is typically set by the Address Update Admin utility. Resides in app.cfg. |
| NCOA/REPOSTYPE | Specifies the repository type | Required | Valid values are: 0 (Guess), 1 (ODBC), 2 (DFTK). If the value is 0, the node attempts to determine the type from the connection string. This is typically set by the Address Update Admin utility. Resides in app.cfg. |
| NCOA/DFAV_CACHE_SIZE | Set verify cache percentage. | Optional | The higher the value the more data is cached, the faster the processing, the more memory used. The default is 0. Resides in macros/ncoa.cfg. |
| NCOA/DFAV_PRELOAD | Set verify preload options. | Optional | Valid values are "ALL" or an empty string. Using "ALL" requires a large amount of memory. Resides in macros/ncoa.cfg. |
| **Pooling** | | | |
| Note that for puddle options, the name of the puddle is placed after 'POOLING/', for example, POOLING/WFEJOB/MAXIMUM_PROCESSES. If no puddle name is specified, it is globally applied to all puddles. Here are a few puddles: WFEJOB - batch jobs on DMServer, WFESVC - Process services on DMSERVER, APISVC - DFAPI services (in the works) | | | |
| POOLING/CHILD_MAXIMUM_LAUNCHES | Throttling for launches | Optional | When specified, the number of concurrent child process launches is limited by this value. If the current child launch request exceeds the specified value, the launch will wait until the number of launching processes is below the specified value. If zero or not specified, there is no limit of concurrent child launches. |
| POOLING/MAXIMUM_PROCESSES | Maximum number of concurrent pooled processes | Optional | If 0, the number of concurrent pooled processes is unlimited. Default is unlimited (previously was 10). If POOLING/GET_PROCESS_TIMEOUT is set, it will wait for that amount of time to get a new process if it needs one. This is a puddle option. |
| POOLING/GET_PROCESS_TIMEOUT | Acquire process timeout | Optional | Default is no timeout. Specifies the length of time, in seconds, the process requester should wait for a process to become available. If zero, the requester will wait indefinitely. The acquire process timeout is in terms of the acquisition of a process and the process pooling handshaking. It does not consider the time required by the requester to complete application level initialization. This is a puddle option. |

| Option | Purpose | Source | Notes |
|---|---|---|---|
| POOLING/IDLE_TIMEOUT | Idle process timeout | Optional | Default is 0. Specifies the length of time, in seconds, a process remains idle before it is terminated. If zero, idle processes are not terminated. This is a puddle option. |
| POOLING/MAXIMUM_ERRORS | Maximum number of pooled process errors before process is terminated | Optional | Default is 0 (never terminate it). This controls how many times a process can fail (when it is reused for something else) before it is terminated. This is a puddle option. |
| POOLING/MAXIMUM_USE | Maximum number of pooled process uses before process is terminated. | Optional | Default is 0 (unlimited). The maximum number of times a pooled process may be used. After the pooled process has been used the specified number of times, it is terminated. This is a puddle option. |
| **Process Flow** | | | |
| WFE/CANCEL_TIMEOUT | Amount of time to give remote processes to cancel in milliseconds | Optional | When user hits cancel the amount of time to wait for remote nodes to exit gracefully before killing them. |
| WFE/MAX_EMBED_COUNT | Maximum embedding level | Optional | Defaults to 100. Maximum level of embedding for process nodes. Helps prevent runaway recursion when recursively embedding jobs. |
| WFE/MAX_EMBED_COUNT | Maximum threads | Optional | Default unlimited. Allows the user to allocate a number of threads for running process nodes. Each thread runs a node (even when running out of process). If this count is reached, nodes scheduled for execution will queue up. |
| WFE/ENGINE_THREAD_LIMIT | Specifies the thread pool limits for the workflow engine. | Optional | Use this setting to limit the number of engine threads. The default is 0 meaning unbounded, which will defer to the system for the thread pool limits. The optimal setting is the number of processors + 1. |
| WFE/STATUS_FREQUENCY | How frequently to update status | Optional | This is milliseconds, the default is 250. This is how long to wait before obtaining status from a remote node. Setting to -1 disables polling for status (which may yield better performance). |
| **Profile** | | | |
| PROF/DEBUG_MODE | Frequency distribution engine debug mode | Optional | Possible values include 0 not debug mode or 1 debug mode: default is not debug mode. The log is located at C:\Documents and |

| Option | Purpose | Source | Notes |
|---|---|---|---|
| | | | Settings\<USERID>\Local Settings\Temp. |
| PROF/PER_TABLE_BYTES | Frequency distribution engine per table bytes | Optional | Any numeric value. Default is -1 (frequency distribution engine default). |
| PROF/LOCK_RETRIES | SQLite repository connection attempts | Optional | Specifies the number of times to retry SQLite repository connection when a connect attempt times out or -1 to retry until a connection is established. Used by explorer too. |
| **QKB** | | | |
| QKB/PATH | Path to QKB | Required by QKB products | Path is set to the default QKB defined in application. |
| QKB/SURFACEALL | Surfaces all parse definitions | Optional | Default is NO. Note that the application start-up creates a blue fusion pool that sets the option for all consumers (Profile, Explorer and Nodes) with the exception of the Expression Engine, which continue to have its own initialization. |
| QKB/COMPATVER | Denotes which compatible version to use when interacting with QKB | Optional | Possible values: dfpower82, dmp21, and dmp22. Default: dmp22. Note that the application start-up creates a blue fusion pool that sets the option for all consumers (Profile, Explorer and Nodes) with the exception of the Expression Engine, which continues to have its own initialization. |
| QKB/ALLOW_INCOMPAT | Allow use of incompatible Quality Knowledge Base definitions | Optional | Default is NO. Note that the application start-up creates a blue fusion pool that sets the option for all consumers (Profile, Explorer and Nodes) with the exception of the Expression Engine, which continues to have its own initialization. |
| QKB/ON_DEMAND | Loads QKB definitions on demand | Optional | Default is YES. Note that the application start-up creates a blue fusion pool that sets the option for all consumers (Profile, Explorer and Nodes) with the exception of the Expression Engine, which continues to have its own initialization. |
| CUSTOMIZE/DISABLE_FILE_NOTIFICATIONS | Temporarily disables notifications | Optional | Read by ui.cfg when Customize starts. When QKB developers make numerous small changes to files in an editor while Customize is open, Customize sends a notification that warns that the file being changed and provides a |

| Option | Purpose | Source | Notes |
|---|---|---|---|
| | | | list of all the definitions that are affected. To temporarily disable these notifications, edit ui.cfg by adding CUSTOMIZE/DISABLE_FILE_NOTIFICATIONS=1. |
| **Architect Client (UI) settings** | | | |
| ARCHITECT/AutoPassThru | Client option to set mappings | Optional | Maintained by client; choices are 0 (target), 1 (Source and Target), and 2 (All). |
| **Architect nodes, etc (Defined in ids.h)** | | | |
| CLUSTER/BYTES | Specifies the bytes use when clustering | Optional | |
| CLUSTER/LOG | Specifies whether a clustering log is needed | Optional | |
| CLUSTER/TEMP | Specifies the cluster temporary path | Optional | |
| FRED/LOG | Specifies whether a FRED log is needed | Optional | |
| JAVA/CLASSPATH | Specifies the Java classpath | Optional | |
| JAVA/DEBUG | | Optional | |
| JAVA/DEBUGPORT | | Optional | |
| VERIFY/USELACS | Enables or disables the LACSLink processing | Optional | Locatable Address Conversion System (LACS). |
| VERIFY/USEELOT | Enables or disables the eLOT processing | Optional | |
| VERIFYINTL/CFG | Verifies the international addresses | Required by international verification | Path maintained by component installation. |
| VERIFYWORLD/CONFIGFILE | Specifies the path to the SetConfig.xml file that is used by the Address Verification (World 2) data job node. Use this option to change the | Required by Address Verification (World 2) data job node | For more information about this file, see the Data Management Studio documentation for the Address Verification (World 2) node. |

| Option | Purpose | Source | Notes |
|---|---|---|---|
| | default location of this file. | | |
| VERIFYWORLD/DB | Specifies the Platon data path | Required for Platon | Path maintained by component installation. |
| VERIFYWORLD/UNLK | Specifies the Platon library universal unlock code | Required for Platon | Path maintained by component installation. |
| WEBSERVICE/CONFIG_FILE | Specifies a user-defined configuration file for the Web Service node and the HTTP Request node. This file can be used to increase the time-out value, for example. | Optional | For more information about the user-defined configuration file, see the FAQ topic: "What Can I Do About Time-Out Errors in Data Jobs with the Web Service Node or the HTTP Request Node?" in the *DataFlux Data Management Studio: User's Guide*. |
| **dfIntelliServer** | | | |
| DFCLIENT/CFG | Used for dfIntelliServer | Required | Maintained by dfIntelliServer installation; typical location is 'C:\Program Files\DataFlux\dfIntelliServer\etc\dfclient.cfg; modify the dfclient.cfg file to point to the server and port. |
| **Repository** | | | |
| REPOS/CREATE_SPEC_PATH | Specifies how to create the repository table or index | Optional | This specification provides a means of configuring the commands to create tables and indexes in the repository. |
| REPOS/FORCE_FILE_BASED | Repository SQLite usage | Optional | If set to true all SQLite access will go through dfsqlite instead of DAC. |
| REPOS/LOCK_RETRIES | Specifies the number of attempts to connect to a SQLite repository | Optional | Number of times to retry SQLite repository connection when a connect attempt times out or -1 to retry until a connection is established. |
| REPOS/TABLE_LIST_PATH | Repository XML table definition | Optional | The directory that should contain XML files for any tables the repository library should add on creation or update. If set we look in here for XML files that will contain repository table definitions, if not set we look in DFEXEC_HOME/etc/reposcreate. |
| | | | |

| Option | Purpose | Source | Notes |
|---|---|---|---|
| **Other** | | | |
| EXPRESS_MAX_STRING_LENGTH | Specifies the maximum size of strings declared in expression nodes | Optional | Default maximum length of any string in this node is 5,242,880 bytes (5MB). This enables specifying a larger value in bytes. If performance issues arise, the suggested setting is 65536 bytes. |
| EXPRESSION/UDFDIR | Specifies where to look for UDF files | Optional | If not specified, UDF files will be sought in installationdir/etc/udf. |
| JAVA/COMMAND | Command used to launch Java | Optional | Default is java. This is the command used to launch the java proxy process. The Java command must be compatible with launching from the command line. Here are some examples:<br>`JAVA/COMMAND = java`<br>`JAVA/COMMAND = java -Djavax.net.ssl.trustStore=C:\Store\jssecacerts`<br>`JAVA/COMMAND = java -Djavax.net.ssl.trustStore="C:\Cert Store\jssecacerts"`<br>`JAVA/COMMAND = "C:\Program Files\Java\jre6\bin\java"` |
| MDM/REPOSITORY_ROOT_FOLDER | | Optional | Name/location of the root folder for mdm within a repository. This one allows the “foundations/master_data” to be overwritten by the end user when putting the contents of [INSTALL_ROOT]/share/mdm into a repository. |
| STEPENG/PROFILEBYNODE | Specifies the performance profiler by node instance. | Use only for design and testing. Do not use in a production environment | When set to Yes, this gives you each node instance and how many milliseconds were spent on each of three operations (prepare, pre-execute, execute), and how many times each was entered. The ID will correspond to the iid field in the XML file for the job, and will include the job name so you can see embedded jobs. To turn on the functionality, go into your configuration files. To profile realtime services, update dfwsvc.cfg. To profile batch jobs, update dfwfproc.cfg. To profile from studio, update ui.cfg. To profile all three, update app.cfg. The results are written to the log under the DF.RTProfiler heading at trace level.<br>An example of the output is: **NX,inner2.ddf,1,0,5**<br>where the values represent action type |

| Option | Purpose | Source | Notes |
|---|---|---|---|
| | | | or operation (either NX - cumulative time spent processing rows, PR - time preparing, or PX - time pre-executing), job name, instance (iid field in the XML file), milliseconds, and entries (the number of times you've entered that code). |
| STEPENG/PROFILEBYTYPE | Specifies the performance profiler by node type. | Use only for design and testing. Do not use in a production environment | When set to Yes, this setting provides you each node type along with how many milliseconds were spent on each of three operations (prepare, pre-execute, execute), and how many times each was entered.<br>To turn on the functionality, go into your configuration files. To profile realtime services, update dfwsvc.cfg. To profile batch jobs, update dfwfproc.cfg. To profile from studio, update ui.cfg. To profile all three, update app.cfg. The results are written to the log under the DF.RTProfiler heading at trace level.<br>An example of the output is: **NX,ARCHITECT_EMBEDDED_JOB,0,5**<br>where the values represent action type or operation (either NX - cumulative time spent processing rows, PR - time preparing, or PX - time pre-executing), node type, milliseconds, and entries (the number of times you've entered that code). |
| SAP_LIBPATH | Specifies the location of SAP RFC libraries on UNIX only. | Optional | These shared libraries are installed to support the SAP Remote Function Call node, a data job node in Data Management Studio. For more information, see "Installing Support for the SAP RFC Node" topic in the DataFlux Data Management Studio Installation and Configuration Guide. |

# Data Access Component Directives

The Data Access Component (DAC) enables you to connect to data using Open Database Connectivity (ODBC) and Threaded Kernel Table Services (TKTS). ODBC database source names (DSNs) are not managed by the DAC, but by the Microsoft ODBC Administrator. TKTS DSNs, however, are managed by the DAC, and TKTS connections are stored in a TKTS DSN directory. The DAC reads a list of settings to determine what actions to take, where to find various things it needs, and how to make the proper connections.

**Note:** The default DAC directives should be satisfactory for most sites. Change these settings only if you have special needs.

There are three types of DAC Configuration Directives:

**Configuration options**: Most of the DAC's settings come from configuration values as specified in app.cfg or other config files that are read in a specific sequence. It is recommended to put these settings in the app.cfg file, but they can be put in macro.cfg or another application-specific configuration file. See Configuration Options for a more comprehensive list of possible configuration values.

**Registry**: The DAC only looks in the registry for DSN-specific items. These are found in either or both HKEY_CURRENT_USER\Software\DataFlux\dac\{version}\{dsn} or HKEY_LOCAL_MACHINE\Software\DataFlux\dac\{version}\{dsn}. "dsn" in this case is a string with the same name as the DSN to be affected.

**Environment variables**: You can identify the environment variables with $NAME in the following table. Typically, the application using the DAC will set environment variables to appropriate locations. On Windows, $DFEXEC_HOME is set to the application's root directory.

| Setting | Description | Windows | UNIX |
|---|---|---|---|
| Command file execution | Specifies a text file with SQL commands (one per line). These commands will run in turn, on any new connection that is made. For example, they can be used to set session settings. This is implemented only for the ODBC driver. | The DAC/SAVEDCONNSYSTEM configuration value may specify the path to the saved connections. The DAC checks for files with the same filename as the DSN and a .sql extension. | The same as Windows. |
| Connection and login timeout | Allows you to specify the time in seconds for a connection timeout and a login timeout. | The string type value is located in USER*dsn_name*\logintimeout (or connecttimeout) where dsn_name is the name of the dsn. The string value is the number of seconds for the time out. If it does not find it, the | Not supported. |

| Setting | Description | Windows | UNIX |
|---|---|---|---|
| | | DAC will look in the SYSTEM*dsn_name*\logintimeout. | |
| DAC logging | Determines whether to create a log file for DAC operations. | This is controlled through the effective logging configuration files, which uses log4sas. The logger name "DAC" should be used. | The same as Windows. |
| DFTK log file | Specifies the log file that interacts with the DFTKSRV layer and is useful only for debugging issues specific to dftksrv. This setting is applicable only if you are running DFTK out of process. | Look for config value DAC/DFTKLOGFILE. | The same as Windows. |
| Disable CEDA | Specifies whether to disable CEDA. This setting is applicable only to tkts connections. | Look for DAC/DFTKDISABLECEDA configuration value, which should specify any non-null value, for example, yes. | The same as Windows. |
| Multiple active results sets (MARS) | Specifies whether to allow a connection to open multiple result sets at the same time. This is applicable only when using the SQL Server Native Client driver on Windows against an SQL Server 2005 (or later). | Look in USER*dsn_name*\mars for dword value of 1 where *dsn_name* is the name of the dsn. If it does not find it, the DAC will look in the SYSTEM*dsn_name*\mars. | Not applicable. |
| Oracle NUMBER(38) handling | If connected to Oracle (only), NUMBER(38) columns is treated as INTEGER by default. To override that functionality, treat them as REAL. | Look in USER\oranum38real for dword value of 1. If it is not found in USER, look in SYSTEM\oranum38real. | Look in $HOME/.dfpower/dsn.cfg for a line dsn_name = oranum38real where dsn_name is the name of the DSN. |
| Read uncommitted | For SQL Server only, specifies that reading data from cursors is allowed to read uncommitted rows. This can alleviate lock ups in SQL Server (applies to DMP version 2.3 and beyond) | Look in USER\dsn_name\readuncommitted for dword value of 1. Where dsn_name is the name of the dsn. After looking in USER look in SYSTEM | Look in $HOME/.dfpower/dsn.cfg for a line dsn_name = readuncommitted where dsn_name is the name of the DSN |
| Run DFTK out of process | Specifies whether to run TKTS out of process, allowing you to perform troubleshooting. | Look for config value DAC/DFTK_PROCESS. This should specify any non-null value, for example, yes. | The same as Windows. |

| Setting | Description | Windows | UNIX |
|---|---|---|---|
| Suffix for CREATE TABLE statements | This allows you to provide a string that is appended to every CREATE TABLE statement. If you include %t in this string, it is substituted with the table name. | Look in USER*dsn_name*\postcreate for a string where dsn_name is the name of the dsn. If it is not found in USER, look in SYSTEM. | Not supported. |
| System saved connection | Specifies where to find system saved connections. | Look in DAC/SAVEDCONNSYSTEM config value for a path. If it is not found, the DAC checks for the DFEXEC_HOME environment variable in the $DFEXEC_HOME\etc\dsn subdirectory, which is in the installation directory. | Look for the same config value. If not found, look in the $DFEXEC_HOME/etc/dsn subdirectory, which is in the installation directory. |
| TK Path | Specifies where TK files are located. This setting is applicable only if you are running Data Factory Took Kit (DFTK) out of process. The dftksrv path and core directory should be specified. | Look for config value DAC/DFTK_PROCESS_PATH. If still not found, set it to $DFEXEC_HOME\bin;$DFEXEC_HOME\bin\core\sasext. | Same as for Windows, but will set it to $DFEXEC_HOME/lib/tkts if it is not found. |
| TKTS DSN directory | Specifies the path where TKTS DSNs are stored in XML files. | Look in config value DAC/DSN. If it is not found, look in $DFEXEC_HOME\etc\dftkdsn\. | Look in the same config value. If it is not found, look in $DFEXEC_HOME/etc/dftkdsn/. |
| TKTS log file | Specifies the log file that is produced by the TKTS layer and is useful for debugging tkts issues. | Look for config value DAC/TKTSLOGFILE. | The same as Windows. |
| TKTS startup sleep | Specifies how much time in seconds to delay between the start of the dfktsrv program and the booting of TK. This setting is applicable only if you are running DFTK out of process. | The DAC checks the registry for a tktssleep value. | Not supported. |
| User saved connection | Specifies where to find user-saved connections. | Look in DAC/SAVEDCONNUSER config value for a path. If not found, look in the user's application settings directory, which is usually in the %APPDATA% directory, in the DataFlux\dac*version* | Look for the same config value. If not found, look in the $HOME/.dfpower/dsn directory. |

| Setting | Description | Windows | UNIX |
|---|---|---|---|
| | | subdirectory. The %APPDATA% location is C:\Users\userid\AppData\ for Windows 7 and C:\Documents and Settings\userid\Application Data\ for Windows XP. | |
| Use braces | Specifies whether to enclose DSN items with braces when they contain reserved characters. Most drivers do not support use of braces; however, this might be an option if problems occur with saving credentials. | The DAC looks in USER\usebraces for dword value of 1. If it is not found there, the DAC looks in SYSTEM\usebraces. | Look in $HOME/.dfpower/dsn.cfg for a line dsn_name = usebraces where dsn_name is the name of the DSN. |

# Logging Options

The following log files are provided for Data Management Studio:

- Studio log
- Platform log
- DAC log
- TKTS log

If enabled, the TKTS log will capture TKTS events for Data Management Studio, such as BASE data sets and data sent across the wire to the DataFlux Federation Server. The TKTS log provides a default set of information that cannot be customized.

The Studio log, the Platform log, and the DAC log are enabled by default. These logs use the SAS 9.2 Logging Facility, a flexible and configurable framework that you can use to collect, categorize, and filter events. Then you can write them to a variety of output devices. The logging facility supports problem diagnosis and resolution, performance and capacity management, and auditing and regulatory compliance.

The logging facility framework categorizes and filters log messages in SAS server and SAS programming environments. It also writes log messages to various output devices. In the server environment, the logging facility logs messages based on predefined message categories such as the following:

- Admin for administrative messages
- App for application messages
- Perf for performance messages

Messages for a category can be written to files, consoles, and other system destinations simultaneously. The logging facility also enables messages to be filtered based on the following thresholds, in order from most inclusive to least inclusive: TRACE, DEBUG, INFO, WARN, ERROR, and FATAL. The configuration files for the Studio log, the Platform log, and the DAC log contain a template for logging options. You need only to un-comment certain lines to turn on the corresponding options. For a full description of these options, see the *SAS Logging: Configuration and Programming Reference*.

The following table specifies the scope, level thresholds, and configuration location for each of these logs:

| Name | Scope | Level Thresholds | Configuration Location |
|---|---|---|---|
| Studio | Data Management Studio events | TRACE, DEBUG, INFO, WARN, ERROR, and FATAL | By default, these are enabled. If you want to turn them off, comment out the BASE/LOGCONFIGPATH line in ui.cfg and in batch.cfg and restart Data Management Studio.<br><br>For configuration, go to *drive:*\Program Files\DataFlux\DMStudio\studio1\etc\platform.log.xml and *drive:*\Program Files\DataFlux\DMStudio\studio1\etc\dfwfproc.log.xml |
| Platform | Data Management Studio engine events | TRACE, DEBUG, INFO, WARN, ERROR, and FATAL | |
| DAC | Data access events | TRACE, DEBUG, INFO, WARN, ERROR, and FATAL | |
| TKTS | TKTS events such as BASE data sets and data sent across the wire to Federation Server | Not configurable | *drive:*\Program Files\DataFlux\DMStudio\studio1\etc\app.cfg |

By default, all Windows XP log files are written to *drive:*\Documents and Settings*USERNAME*\Application Data\DataFlux\DMStudio\[*version*]\Logs. The default location in Windows 7 is drive:\Users\[username]\AppData\Roaming\DataFlux\DMStudio\[version]\Logs.

Also by default, the Studio, Platform and DAC logging are all logged to the platform log file. This log file has a name prefixed with *platform_*. The DAC logging can be redirected to its own log file by uncommenting the appender-ref setting in the logger section for *DF.DAC* and by uncommenting the RollingFileAppender section for *DacFile*. The latter contains the filename pattern that dictates where and what name the DAC log file takes. The Studio logging can be redirected to its own log file by uncommenting the appender-ref setting in

the logger section for *DF.DMStudio* and by uncommenting the RollingFileAppender section for *DMStudioFile*. The latter contains the filename pattern that dictates where and what name the Studio log file takes.

The Studio, Platform and DAC logging all use a logging level of INFO by default. This level is set in the root logger section of the platform.log.xml configuration file. The level value can be changed here to change the default for all three logs. In addition, the logging level for the DAC and Studio logs can be set separately in the *DF.DAC* and *DF.DMStudio* logger sections respectively by uncommenting (and possibly updating) the level values in each.

The platform.log.xml file controls all of Studio, Platform and DAC logging within the Data Management Studio application except for job runs. When you run a job within Data Management Studio, Data Management Studio creates a separate dfwfproc process for that job run. Therefore, to configure logging for jobs run from Data Management Studio, you should modify the dfwfproc.log.xml.

# Installing Add-On Products



## Installing QKBs and Data Packs

A Quality Knowledge Base (QKB) is a collection of files that define rules, patterns, and transformations that are used for data cleansing. Data Packs are collections of third-party enrichment data, such as Loqate international address data. QKBs and DataPacks are not installed as part of the SAS Deployment Wizard process. If you have licensed a QKB or a Data Pack, go to MyDataFlux Portal to download and install these products. Use the Release Notes for your QKBs in order to verify that they are compatible with your version of SAS software.

QKB: http://www.dataflux.com/qkb

DataPacks: http://www.dataflux.com/Resources/DataFlux-Resources/MyDataFlux-Portal/Downloads/Data-Updates.aspx

After you install a QKB, see the "Using Quality Knowledge Bases" chapter in the *DataFlux Data Management Studio: User's Guide* to enable Data Management Studio to access the QKB. After you install a data pack for a feature in Data Management Studio, see the topic for that feature in the *DataFlux Data Management Studio: User's Guide* to enable Data Management Studio to access the data pack.

## Installing Supplemental Language Support

If you plan to use DataFlux Data Management Studio (Studio) for data that includes East Asian languages or right-to-left languages, you must install additional language support. Complete these instructions to install these packages:

1. Click **Start** > **Settings** > **Control Panel**.
2. Double-click **Regional and Language Options**.
3. In the Regional and Language Options dialog, select the **Languages** tab.
4. Under **Supplemental Language Support**, select the check boxes marked, **Install Files for complex script and right-to-left languages (including Thai)** and **Install files for East Asian languages**.
5. The Microsoft Windows installer guides you through the installation of these language packages.

# Installing the Address Update Add-On

The DataFlux® Address Update add-on enables you to use the United States Postal Service (USPS) NCOALink® system to identify and update customer address information. To install the required resources for this add-on, see the *Address Update Add-On to Data Management Studio Quick Start Guide*. This document is provided to customers who license the Address Update add-on.

# Installing Support for the SAP RFC Node

If you open a data job in Data Management Studio, the **SAP Remote Function Call** node is available in the **Data Integration** folder. This node enables you to read and write data using SAP RFC and Business Application Programming Interfaces (BAPI) provided by SAP. This functionality would be useful to

- read information from SAP that is not readily available in tables
- use the functionality that SAP exposes as BAPIs
- load data to SAP by using SAP-provided functions that are a standard way of updating objects
- trigger an upload to SAP by using functions that take, for example, an external file input.

In order to use the SAP Remote Function Call node, you must install SAP libraries (DLLs) on the computer where the job that contains the node will be executed. For example, if you will use Data Management Studio to execute a job with an **SAP Remote Function Call** node, then you must install the SAP libraries on the computer where Data Management Studio is running. If you will use Data Management Server to execute a job with an SAP Remote Function Call node, then you must install the SAP libraries on the computer where Data Management Server is running.

Data Management Studio is a 32-bit application and so requires 32-bit SAP libraries. See the SAP documentation for details about installing these libraries.

Data Management Servers can be either 32-bit or 64-bit. See the SAP documentation for details about installing the appropriate libraries.

Data Management Studio is a Windows application, but Data Management Servers can run on UNIX. For UNIX systems only, you can set the SAP_LIBPATH option in the app.cfg file to point to the SAP DLLs.

# Technical Support



# Frequently Asked Questions (FAQ)

The following questions and answers are designed to assist you when working with Data Management Studio.



## System Requirements

### What are the recommended system requirements for Data Management Studio?

The recommended system requirements for Data Management Studio can be found in System Requirements.

## Supported Platforms

### On what platforms can I install Data Management Studio?

Data Management Studio can be installed on many different platforms. For the full list of supported platforms, see System Requirements.

## Data Storage

### From what databases can DataFlux Data Management Studio access data?

Data Management Studio can access data from any of the databases listed in the Data Storage topic.

## Repository Storage

### What databases are supported for Data Management Platform repositories?

Data Management Studio can access data from any of the databases listed in the Database Storage for Repositories topic.

# Logging Options

### What options do I have for logging?

See the Logging Options section for the scope, level thresholds, and configuration location for the available logs.

# Appendixes



# Legal Notices





## DataFlux Legal Statements

### Apache Portable Runtime License Disclosure



Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

### Apache / Xerces Copyright Disclosure

The Apache Software License, Version 3.1



Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

3. The end-user documentation included with the redistribution, if any, must include the following acknowledgment:

   "This product includes software developed by the Apache Software Foundation (http://www.apache.org)."

   Alternately, this acknowledgment may appear in the software itself, if and wherever such third-party acknowledgments normally appear.

4. The names "Xerces" and "Apache Software Foundation" must not be used to endorse or promote products derived from this software without prior written permission. For written permission, please contact apache@apache.org.

5. Products derived from this software may not be called "Apache", nor may "Apache" appear in their name, without prior written permission of the Apache Software Foundation.

THIS SOFTWARE IS PROVIDED "AS IS'' AND ANY EXPRESSED OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE APACHE SOFTWARE FOUNDATION OR ITS CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

This software consists of voluntary contributions made by many individuals on behalf of the Apache Software Foundation and was originally based on software copyright (c) 1999, International Business Machines, Inc., http://www.ibm.com. For more information on the Apache Software Foundation, please see http://www.apache.org.

### Boost Software License Disclosure

Boost Software License - Version 1.0 - August 17, 2003

Permission is hereby granted, free of charge, to any person or organization obtaining a copy of the software and accompanying documentation covered by this license (the "Software") to use, reproduce, display, distribute, execute, and transmit the Software, and to prepare derivative works of the Software, and to permit third-parties to whom the Software is furnished to do so, all subject to the following:

The copyright notices in the Software and this entire statement, including the above license grant, this restriction and the following disclaimer, must be included in all copies of the Software, in whole or in part, and all derivative works of the Software, unless such copies or derivative works are solely in the form of machine-executable object code generated by a source language processor.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. IN NO EVENT SHALL THE COPYRIGHT HOLDERS OR ANYONE DISTRIBUTING THE SOFTWARE BE LIABLE FOR ANY DAMAGES OR OTHER LIABILITY, WHETHER IN CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

### DataDirect Copyright Disclosure

Portions of this software are copyrighted by DataDirect Technologies Corp., 1991 - 2008.

### Expat Copyright Disclosure

Part of the software embedded in this product is Expat software.

Copyright © 1998, 1999, 2000 Thai Open Source Software Center Ltd.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

### gSOAP Copyright Disclosure

Part of the software embedded in this product is gSOAP software.

Portions created by gSOAP are Copyright © 2001-2004 Robert A. van Engelen, Genivia inc. All Rights Reserved.

THE SOFTWARE IN THIS PRODUCT WAS IN PART PROVIDED BY GENIVIA INC AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY

DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

## IBM Copyright Disclosure

ICU License - ICU 1.8.1 and later [used in DataFlux Data Management Platform]

COPYRIGHT AND PERMISSION NOTICE

Copyright © 1995-2005 International Business Machines Corporation and others. All Rights Reserved.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, provided that the above copyright notice(s) and this permission notice appear in all copies of the Software and that both the above copyright notice(s) and this permission notice appear in supporting documentation.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR HOLDERS INCLUDED IN THIS NOTICE BE LIABLE FOR ANY CLAIM, OR ANY SPECIAL INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

Except as contained in this notice, the name of a copyright holder shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization of the copyright holder.

## Microsoft Copyright Disclosure

Microsoft®, Windows, NT, SQL Server, and Access, are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

## Oracle Copyright Disclosure

Oracle, JD Edwards, PeopleSoft, and Siebel are registered trademarks of Oracle Corporation and/or its affiliates.

## PCRE Copyright Disclosure

A modified version of the open source software PCRE library package, written by Philip Hazel and copyrighted by the University of Cambridge, England, has been used by DataFlux for regular expression support. More information on this library can be found at: ftp://ftp.csx.cam.ac.uk/pub/software/programming/pcre/.

Copyright © 1997-2005 University of Cambridge. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of the University of Cambridge nor the name of Google Inc. nor the names of their contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY,

WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

### Red Hat Copyright Disclosure

Red Hat® Enterprise Linux®, and Red Hat Fedora™ are registered trademarks of Red Hat, Inc. in the United States and other countries.

### SAS Copyright Disclosure

Portions of this software and documentation are copyrighted by SAS® Institute Inc., Cary, NC, USA, 2009. All Rights Reserved.

### SQLite Copyright Disclosure

The original author of SQLite has dedicated the code to the public domain. Anyone is free to copy, modify, publish, use, compile, sell, or distribute the original SQLite code, either in source code form or as a compiled binary, for any purpose, commercial or non-commercial, and by any means.

### Sun Microsystems Copyright Disclosure

Java™ is a trademark of Sun Microsystems, Inc. in the U.S. or other countries.

### USPS Copyright Disclosure

National ZIP®, ZIP+4®, Delivery Point Barcode Information, DPV, RDI, and NCOA$^{Link}$®. © United States Postal Service 2005. ZIP Code® and ZIP+4® are registered trademarks of the U.S. Postal Service.

DataFlux is a non-exclusive interface distributor of the United States Postal Service and holds a non-exclusive license from the United States Postal Service to publish and sell USPS CASS, DPV, and RDI information. This information is confidential and proprietary to the United States Postal Service. The price of these products is neither established, controlled, or approved by the United States Postal Service.

### VMware

VMware® virtual environment provided those products faithfully replicate the native hardware and provided the native hardware is one supported in the applicable DataFlux product documentation. All DataFlux technical support is provided under the terms of a written license agreement signed by the DataFlux customer.

The VMware virtual environment may affect certain functions in DataFlux products (for example, sizing and recommendations), and it may not be possible to fix all problems.

If DataFlux believes the virtualization layer is the root cause of an incident; the customer will be directed to contact the appropriate VMware support provider to resolve the VMware issue and DataFlux shall have no further obligation for the issue.

## Solutions and Accelerators Legal Statements

Components of DataFlux Solutions and Accelerators may be licensed from other organizations or open source foundations.

### Apache

This product may contain software technology licensed from Apache.

Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at: http://www.apache.org/licenses/LICENSE-2.0.

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.

See the License for the specific language governing permissions and limitations under the License.

### Creative Commons Attribution

This product may include icons created by Mark James http://www.famfamfam.com/lab/icons/silk/ and licensed under a Creative Commons Attribution 2.5 License: http://creativecommons.org/licenses/by/2.5/.

### Degrafa

This product may include software technology from Degrafa (Declarative Graphics Framework) licensed under the MIT License a copy of which can be found here: http://www.opensource.org/licenses/mit-license.php.

Copyright © 2008-2010 Degrafa. All rights reserved.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

### Google Web Toolkit

This product may include Google Web Toolkit software developed by Google and licensed under the Apache License 2.0.

### JDOM Project

This product may include software developed by the JDOM Project (http://www.jdom.org/).

### OpenSymphony

This product may include software technology from OpenSymphony. A copy of this license can be found here: http://www.opensymphony.com/osworkflow/license.action. It is derived from and fully compatible with the Apache license that can be found here: http://www.apache.org/licenses/.

### Sun Microsystems

This product may include software copyrighted by Sun Microsystems, jaxrpc.jar and saaj.jar, whose use and distribution is subject to the Sun Binary code license.

This product may include Java Software technologies developed by Sun Microsystems,Inc. and licensed to Doug Lea.

The Java Software technologies are copyright © 1994-2000 Sun Microsystems, Inc. All rights reserved.

This software is provided "AS IS," without a warranty of any kind. ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE HEREBY EXCLUDED. DATAFLUX CORPORATION LLC, SUN MICROSYSTEMS, INC. AND THEIR RESPECTIVE LICENSORS SHALL NOT BE LIABLE FOR ANY DAMAGES SUFFERED BY LICENSEE AS A RESULT OF USING, MODIFYING OR DISTRIBUTING THE SOFTWARE OR ITS DERIVATIVES. IN NO EVENT WILL SUN MICROSYSTEMS, INC. OR ITS LICENSORS BE LIABLE FOR ANY LOST REVENUE, PROFIT OR DATA, OR FOR DIRECT, INDIRECT, SPECIAL, CONSEQUENTIAL, INCIDENTAL OR PUNITIVE DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF THE USE OF OR INABILITY TO USE SOFTWARE, EVEN IF SUN MICROSYSTEMS, INC. HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

### Java Toolkit

This product includes the Web Services Description Language for Java Toolkit 1.5.1 (WSDL4J). The WSDL4J binary code is located in the file wsdl4j.jar.

Use of WSDL4J is governed by the terms and conditions of the Common Public License Version 1.0 (CPL). A copy of the CPL can be found here at http://www.opensource.org/licenses/cpl1.0.php.

# Glossary

## A

**Access Control Entry**

An Access Control Entry (ACE) is an entry of user information made to the Access Control Lists (ACLs) which is used to secure access to individual DataFlux Integration Server (DIS) objects.

**Access Control Lists**

Access Control Lists (ACLs) are used to secure access to individual DataFlux Integration Server (DIS) objects.

**address verification**

Address verification (validation) is the process of comparing a physical address to a reference database of known physical addresses so the original address can be standardized and corrected according to postal authority standards.

**AIC**

Analyze, Improve, Control (AIC) - DataFlux enables organizations to analyze, improve, and control their data from a single data quality integration platform. DataFlux tools and approaches can help you build a comprehensive set of business rules that can create a unified view of your enterprise data and enhance the effectiveness of CDI, CRM, ERP, legacy data migration, or compliance initiatives.

**AMAS**

Address Matching Approval System (AMAS) is the program the Australia Post administers to certify address verification software.

**API**

Application Programming Interface (API) is a set of software protocols, routines, and/or tools used when building software applications.

**APO**

Army/Air Force post office (APO) is an indication for the USPS.

**Architect Job Templates**

Data Management Studio can be used to modify and build work flows called jobs. These jobs can be delivered as templates that can be fleshed out by consultants or other IT professionals. Many job templates will be designed and delivered with the solution to accommodate such things as address verification, merging, assigning IDs, standardizing data, and so on.

**ASCII**

ASCII (American Standard Code for Information Interchange) is a character set based on the English alphabet

## B

**basic category**

A basic category is a category that represents a single word. Basic categories are the basic building blocks of Grammar rules. Every basic category in a Grammar corresponds to a category in an ordered word list. For this reason, you should design Grammar rules in parallel with word-analysis logic.

**batch processing**

The application of data management routines to data source records in what are often very large groups, usually in processes that require no manual user intervention. Contrast with real-time processing.

**business functions**

These are expressions which are written in a generic manner so they can be reused from multiple rules or applications.

**business rule**

A conditional statement that tells a system running a business process how to react to a particular situation.

## C

**case definition**

A set of logic used to accurately change the case of an input value, accounting for unique values that need to be case sensitive, such as abbreviations and business names.

**CASS**

Coding Accuracy Support System (CASS) is the program the United States Postal Service (USPS) administers to certify address verification software.

**CBSA**

Census Bureau Statistical Areas (CBSA)

**CEDA**

Cross-Environment Data Access (CEDA)

**census string**

The census string is a US Census Bureau designation for the boundary area in which the centroid exists. The census string contains state, county, and other census-type information.

**centroid**

A centroid is the approximate mathematical center of the ZIP or ZIP+4 boundary.

**checks**

These are built-in checks (expressions) that provide a template to the user to build common standard expressions.

**chop table**

A proprietary file type used by DataFlux as a lex table to separate characters in a subject value into more usable segments.

**CMRA**

US Commercial Mail Receiving Agency (CMRA)

**CMSA**

Consolidated Metropolitan Statistical Areas (CMSA)

**Comments**

Comments are text within a code segment that is not executed. Comments can be either C-style (starts with /* and ends with */) or C++ style (starts with // and continues to the end of a line).

**Core Fields**

Default logic to handle data such as name and address, which inform the identity management process.

**CPC**

Canadian Post Certification (CPC) is the SERP program administered by the Canadian Post. This is similar to the CASS certification administered by the USPS.

**CRM**

Customer Relationship Management (CRM)

**custom metrics**

Custom metrics may be used when the standard metrics do not contain the rules you need to accomplish the desired results.

## D

**dashboard**

The dashboard is a Web-based view of the task grid and graphs in the Monitor Viewer.

**data profiling**

A discovery process that uncovers potential problem areas in large amounts of structured data.

**data type**

Not used in the sense of a database data type ("varchar" for instance) but used to describe sets of data values that follow certain rules and conventions. "Name" and "Address" are two examples of data types.

**database**

A collection of tables containing data that can be accessed easily by a computer system.

**definition**

An algorithm available to a DataFlux application.

**derived category**

A derived category is a category composed of one or more other categories. The makeup of a derived category is described using rules.

**dfIntelliServer**

dfIntelliServer provides a real-time or transactional mechanism for communicating with the MCRD through the Architect API. dfIntelliServer has several client libraries (including a Web services client) that can be called from a number of different applications in many different computing environments. dfIntelliServer allows one at a time queries and modifications to the MCRD. dfIntelliServer allows organizations to access Architect jobs through an API that can accept one group of data elements at a time rather than a complete table. This functionality takes advantage of the power of encapsulation of discreet chunks of work in Architect, so a programmer need only make one call to the client API to perform a related set of activities.

**DPV**

Delivery Point Validation (DPV) specifies if the given address is a confirmed delivery point as opposed to being within a valid range of house numbers on the street.

**DSN**

Data Source Name (DSN)

## E

**EEL**

Expression Engine Language (EEL)

**ERP**

Enterprise Resource Planning (ERP)

**ETL**

Extraction, Transformation, and Loading

**event**

An event represents an action which should be taken when a rule fails. Actions can include sending email messages, storing the offending row in the repository, or executing an external process.

**Expression**

This is the DataFlux syntax used in the Business Rule Manager to build business rules.

## F

**field**

Also known as a "variable" or a "column," a single piece of data in a database table. Database tables can have many fields. The user defines the fields. Each field has a unique identifier in the repository. From a data monitoring standpoint, the fields are not tied to any specific database or table but are bound at the time of execution to the current data set or row.

**field set**

A field set is a collection of fields that belong together. These usually represent a table of data and are used to aid in building rules and viewing results.

**FIPS**

Federal Information Processing Standards (FIPS) - A 5-digit number assigned to each county in the U.S. by the Census Bureau. The first 2 digits are the state code, and the last 3 digits are the county number.

**FPO**

Fleet post office (FPO) indication for USPS used for military personnel.

## G

**gender analysis**

An algorithm that can determine the gender of persons by their names.

**gender definition**

A set of logic used to determine the probable gender of a name or identity-type input string.

**grammar**

A proprietary file type used to store hierarchical patterns pertinent to a specific subject area.

**group rule**

A group rule evaluates and applies all rules to groups of data (for example, data grouped by state and the rules evaluated for each state).

## H

**historical metrics**

A historical metric is available when a business rule is run a second time under the same report name. You can view and compare the last two reports.

## I

**identification analysis**

An algorithm that can determine from a known set of options what type of data is represented by a particular subject value.

**identification definition**

A set of logic used to identify an input string as a member of a redefined or user-defined value group or category.

**inputs**

Input fields are the fields where you apply the checks specified in the Rule Manager. This list includes all the fields you have defined in the Business Rule Manager, including the Output fields from custom metrics and any grouped by field.

## J

**job**

The saved configuration settings for a particular task in a Data Management Studio application. You can run jobs interactively or combine them with other jobs and schedule the set of jobs to run on a particular date or time.

## L

**LACS**

US Locatable Address Conversion Service (LACS) is a product/system in a different USPS product line that allows mailers to identify and convert a rural route address to a "city-style" address.

**locale**

The country of origin based on an address or country code.

**locale guessing**

A process that attempts to identify the country of origin of a particular piece of data based on an address, country code, or other field.

## M

**match**

The process of identifying data strings that can be different representations of the same semantic information. For example, the strings Mr. Bob Brauer, Robert J., and Brauer can be considered to match each other.

**match cluster**

A set of records grouped together based on some commonality. Cluster IDs are numeric values used to refer to these clusters. You can append cluster IDs to records in a database to document matches.

**match codes**

The end result of passing data through a match definition. A normalized, encrypted string that represents portions of a data string that are considered to be significant with regard to the semantic identity of the data. Two data strings are said to "match" if the same match code is generated for each.

**match definition**

A set of logic used to generate a match code for a data string of a specific data type.

**match value**

A string representing the value of a single token after match processing.

**MCD**

Minor Civil Division (MCD)

**MDM**

Master Data Management (MDM) focuses on master data shared by several different systems and groups.

**merge**

The process of joining records and eliminating duplicate records from a table based on user-specified conditions and rules.

**metadata**

Information that describes the properties of data, for example when was last accessed or the size of the data value.

**micropolitan**

This term is used in US Census data and refers to a population area including a city with 10,000 to 50,000 residents and surrounding areas.

**MSA**

Metropolitan Statistical Areas (MSA) - The MSA code assigned by the Office of Management and Budget. Use this code as an index key in the MSA file.

## N

**namespace**

A namespace is a unique container created to hold a logical grouping of identifiers.

## O

**Object**

An object is anything that can be stored in the Data Management Studio Navigator and accessed by the Data Management Studio applications.

**objects**

Objects are individual jobs and services.

**ODBC**

Open Database Connectivity (ODBC) - an open standard application programming interface (API) for accessing databases.

**OFAC**

Office of Foreign Assets Control (OFAC) - Federal regulations related to the Patriot Act.

**OLAP**

Online Analytical Processing (OLAP)

**organization**

A company, university, or other type of institution. For example: IBM Corporation, University of Connecticut, or St. Joseph's Hospital

**outputs**

The output field is the field(s) used to apply the rule in the custom metric. Set your output field to serve as the field where the results from your custom metric are collected.

## P

**parse**

The process of dividing a data string into a set of token values. For example: Mr. Bob Brauer, Mr. = Prefix, Bob = Given, Brauer = Family

**parse definition**

A name for a context-specific parsing algorithm. A parse definition determines the names and contents of the sub-strings that will hold the results of a parse operation.

**pattern analysis definition**

A regular expression library that forms the basis of a pattern recognition algorithm.

**phonetics**

An algorithm applied to a data string to reduce it to a value that will match other data strings with similar pronunciations.

**PMB**

A private mailbox (PMB) is categorized as a mailbox located at a mail center other than the post office or home.

**PMSA**

Principal Metropolitan Statistical Areas (PMSA)

**Primary Key**

Primary key is a unique identifier assigned to a database field. Social Security Numbers or a ISBNs are examples of possible primary keys.

## Q

**QAS**

QuickAddress Software (QAS)

**QKB**

The Quality Knowledge Base (QKB) is a collection of files and configuration settings that contain all DataFlux data management algorithms. The QKB is directly editable using Data Management Studio.

**Quality Knowledge Base Locales**

The Quality Knowledge Base (QKB) locales contain the files, file relationships, and metadata needed to correctly parse, match, standardize, and otherwise process data.

## R

**RDBMS**

Relational Database Management System (RDBMS) allows you to access data in a database in unique ways, such as adding tables and records, and joining tables.

**RDI**

Residential Delivery Indicator (RDI)

**real-time processing**

Processing a record or data one piece at a time as it enters a computer system, for financial transactions, for example. Contrast with batch processing.

**record**

Also called a "row" or "observation," one complete set of fields in a database table.

**regular expression**

A mini-language composed of symbols and operators that enables you to express how a computer application should search for a specified pattern in text. A pattern may then be replaced with another pattern, also described using the regular expression language.

**repository**

A Data Management Studio repository is a hierarchical data storage mechanism.

**row rule**

A row rule evaluates every row of data passed into the Monitoring node.

**RP**

Software Evaluation and Recognition Program is a program the Canada Post administers to certify address verification software.

**rule**

A single rule can be either a row level rule or a data set level rule. A row level rule is applied to each row which enters the system while a data set level rule is applied to an entire data set or a portion of a data set.

**rule set**

A rule set is a set of one or more rules which are applied together as a group. Use a rule set when you find you are using a few rules together frequently.

## S

**SDK**

Software Development Kit (SDK)

**sensitivity**

Regarding matching procedures, sensitivity refers to the relative tightness or looseness of the expected match results. A higher sensitivity indicates you want the values in your match results to be very similar to each other. A lower sensitivity setting indicates that you would like the match results to be "fuzzier" in nature.

**SERP**

The Software Evaluation and Recognition Program (SERP) is a program the Canadian Post administers to certify address verification software.

**Service Oriented Architecture**

Service Oriented Architecture (SOA) - All of the interaction with the master customer reference database is through a service-oriented architecture that enables any system to talk to the customer database and request or update information.

**set rule**

A set rule evaluates and applies rules to all of the input data completely (for example, it will evaluate all 1000 rows of data as a set).

**SQL**

Structured Query Language (SQL) is a language used to request information from database systems.

**standard metrics**

Standard metrics are pre-defined rules (expressions) set in Data Management Studio. Most of the time, this is enough to achieve the results for your job.

**standardization definition**

A set of logic used to standardize a string.

**standardization scheme**

A collection of transformation rules that typically apply to one subject area, like company name standardization or province code standardization.

**standardize**

The process of transforming a data string so each of the string's token values conforms to a preferred standard representation: IBM Corporation = IBM CORP; Mister Bob Brauer, Junior = MR BOB BRAUER JR.

**Statement of Accuracy**

Statement of Accuracy (SoA) is the form used for Canadian Post Certification (CPC) standards.

## T

**table**

A table is a collection of records in a database.

**tasks**

Tasks contain the rules and the events that go with your individual rule. Tasks associate alert events with a rule that are triggered after a rule fails.

**token**

Used by DataFlux to designate the output strings of a parse process. The output string of a parse process. A word or atomic group of words with semantic meaning in a data string. A set of expected tokens is defined for each data type.

## U

**Unicode**

An industry standard used to allow text and symbols from languages around the world.

**unified**

This is the version of the repository you are using. The term "unified" means the repository contains data for Data Management Studio Profile reports, Business Rules, and Data Monitoring results.

**URI**

Uniform Resource Identifier (URI) is a string of characters identifying a resource or file path.

**USPS**

United States Postal Service (USPS) provides postal services in the United States. The USPS offers address verification and standardization tools.

## V

**vocabulary**

A proprietary file type used for categorizing data look-ups pertinent to a specific subject area.