IDIOM
PURE DECISIONING

Data Masking and Data Migration

Product Description

March 2014

Contents

INTRODUCTION

The IDIOM Decision Manager Workbench [DMW] was originally designed as a large scale, high performance audit tool. Additional capabilities have extended this into the realms of transaction level regression testing, analysis, and reporting; and of transaction level what-if analysis and simulation. A brochure for the product can be found here.

DMW is in use with some of Australia's largest financial organisations.

This document outlines two powerful new capabilities that are now also readily achieved with DMW. These are data masking and data migration. Each is outlined in a separate section in this document.

For background information on the IDIOM Decision Manager Workbench please also see the product page on our website.

Note: reference to 'customer' in this document means an organisation who is a DMW customer of IDIOM.

IDIOM Contact Details

Mark Norton | Director | Idiom Limited |
Office +64 9 630 8950 | Mob +64 21 434669 | After Hrs +64 9 817 7165 |
Aust. Free Call 1 800 049 004
1-8 93 Dominion Rd. Mt Eden Auckland 1024 | PO Box 60101 Titirangi Auckland 0642 | New Zealand

Email mark.norton@idiomsoftware.com | Skype Mark.Norton |

DATA MASKING

Introduction

Data masking is the process of reading database records that contain personally identifying data and copying them to a new instance of the database with all personally identifying data removed and replaced by equally valid but otherwise false data.

The database records in question are typically root data records for the database, for instance a member in a member database, or a client in a client database, and their dependent records (we will refer to these records as 'Person' records in this document regardless of their intent or purpose). The full extent of the personally identifiable data is usually spread across the root and the dependent records. With IDIOM data masking, the root and all of these dependent records are collated into a single Person record instance within the masking process and managed as a single data unit.

It is also common for relationships to exist between the Persons as described above – for instance, relationships that link spouses, or beneficiaries, etc. – and for some of this related party information to be duplicated onto one or more of their related party Person records. Where this copied data is linked via a foreign key (whether database defined or not), it will be masked in a consistent way across the entire database; third party data that is not connected via a foreign key is randomly masked.

Data masking as described uses the standard IDIOM Decision Manager Workbench platform [DMW] and is an inherent part of the DMW platform, subject to the existence within DMW of appropriate masking data for the masking jurisdiction.

This section describes how data masking is implemented using the IDIOM Decision Manager Workbench.

Data Masking – What it Does

On a scheduled and/or on demand basis there is a need to copy live production data to other databases for various purposes (databases for testing, simulation, training, etc.). Modern day privacy requirements often require that all or most personally identifying data be removed from the copied data for these databases. However, for these databases to be useful for all of the intended purposes, the replacement data must be valid in all respects so that the masked data will respond to programs in the same manner as the original data.

With the IDIOM data masking solution, this equivalence of data is sufficient to allow 'trouble-shooting' of production databases even though the personal details have been replaced. In these cases it may be necessary to locate a specific production data record in the masked copy, even though the record itself is masked and not identifiable. To achieve this, DMW provides a special function for two appropriately authorized people working together to identify the equivalent masked record for any given production data record.

The masking data itself, and how it is applied to production records, are both 'extensible' - that is, they are not visible to the DMW application code or its database and so can be extended on a customer by customer basis. DMW currently includes a data file of three million masking identities complete with

names, addresses (incl. valid postcodes), telephone numbers (incl. valid State area codes), tax numbers (TFN), bank account numbers, and company numbers (ABN), all of which are valid for the Australian jurisdiction, but which are otherwise false. Other jurisdiction data, including NZ, can be supplied for customers in other jurisdictions as required.

Third party relationship data (for example, spouse, beneficiary, etc) remains consistent across both the third party record and any data copied onto its related records, so that all data for relationships that cross Person records is consistent across the full extent of the database – that is, any reference to Person X will always be masked with the same data, regardless of where Person X's data is located.

How it Works

Pre-Processing Activities

A target masked database must be provided by the customer as a pre-cursor to the masking process; the access credentials for this database will need to be configured into DMW. A SQL Script can then be supplied to DMW and run as a pre-cursor activity to pre-set this database into any required state – for instance, to generate empty tables, strip down existing tables, and/or remove constraints. This script can also be used to copy all 'reference data' tables to the target database with their keys and contents unchanged. These are all tables that do not contain sensitive information and therefore require no masking.

Post-Processing Activities

Once the masked data has been inserted, a post-cursor SQL Script can be executed to set database triggers, constraints, auto numbering attributes and indexes etc. that may need to be reapplied in order for the applications that use these databases to function as normal.

Masking Data

Three million personal identity records have been pre-loaded into DMW for use in data masking activities. The current data includes the following, all valid for the Australian jurisdiction.

- First Name
- Last name
- Company Name
- Address
- City
- State
- Post Code
- Phone number X 2
- Email address
- Web site if a company
- New masked Id

- ABN
- TFN
- Bank Accounts

This masking data is held in a permanent table within DMW, and will in turn be used to build the operational 'Masking' table that is keyed by existing Person record Ids. All of the above masking data is held as encrypted XML within a single column in both tables mentioned above (although the encryption key for the two tables is different – see the section on Security for an explanation of this).

Masking Person Related Data

The IDIOM Data Mapper (a component of DMW) joins all dependent records for each Person (being a client or a member for instance) into an XML 'Transaction Document[1]' per Person – this document contains all data for the subject Person (including any foreign keys for related Person(s) identified on these records). At this point the original Person identifier(s) are in the clear. The first step in the DMW Data Masking process is to use a decision model to locate any Person identifiers within this record that need to be masked (i.e. the primary and related Person identifiers) and extract them into a standalone 'Masking Document' [another XML document] – this new document is a working document that is only used internally within the Data Masking process. A temporary zero based sequence number is generated by the decision model and used to replace the original Person identifiers in the Transaction Document itself. This new surrogate identifier will be used as the identifier for the subsequent masking activity. This approach is taken so that the original and masked identifiers are never visible at the same time to any decision models. The DMW runtime will use each of the retrieved Person identifiers in the Masking Document to access the DMW Masking table, retrieve the masking data for that Person identifier, decrypt it, and use it to a) substitute the new masking identifier for the original Person identifier; and b) supply a full set of replacement personal data for this (masked) Person. This new masking information, which no longer has the original Person identifier, is then available to subsequent decision models to use for masking purposes, using the surrogate sequential identifier to indirectly match the new masking data with the original person.

Note that for security reasons this injection of masking information occurs immediately preceding the execution of the mapping decision model and replaces the original person key so that the original and masking keys are never available together.

Decision Models

There are two important decision models used in the process. The masking decision models are built bespoke per customer, to allow full consideration of all of the nuances in each customer's database and approach. For instance, one masked database might leave postcodes intact if it is used for geographic simulation, another might want them replaced.

[1] A 'Transaction Document' is a proper DMW term for the primary document being processed through a DMW process.

Pre-Masking Decision Model

The pre-masking decision model manufactures the Masking Document, creating one element for each Person identifier to be masked in this 'Transaction Document', keyed by a derived sequence number. The Person identifiers include:

- The original Person record (the ID is already known)
- All related persons referenced by Person identifier foreign keys (e.g. spouse etc)

Only the Person identifier(s) will be written to the Masking Document in this process. It is the responsibility of the DMW runtime to replace this Person identifier with the masked identifier, and to provide the additional masking information from the Masking table.

A few random spare identities (i.e. not Person ID related) for use in masking personal details that do not have foreign keys are also added by the DMW runtime. These will not have a Person identifier on either the original or updated elements.

This ensures that all of the masking data is available within one XML document to enable complete data masking to occur for each Person record.

The Masking Decision Model

The joining of all Person related database tables into a single Person 'Transaction Document' allows a single decision model to substitute all masking data in one step. The Masking Decision Model walks through the entire 'Person' Transaction Document (which may be substantial – perhaps hundreds of thousands of nodes) selectively replacing all personally identifying data under the control of its customer defined rules.

The Complete DMW Masking Process

The full masking process usually includes a series of linked Jobs – a Job is a proper DMW term and is a standalone unit of work that allows DMW to mix parallel and sequential processing elements into a single overall process. The following is not a rigid process pattern, and serves as an example only.

Primary Job (Target Database Setup)

- Uses a single DMW runtime.
- Executes a SQL (DDL) script to set up the target database and optionally copy all of the reference data.
- Once completed it launches the secondary dependent job(s) which can now run in parallel.

Secondary dependent Job (Data Masking)

- Can use multiple DMW runtimes to process in multiple parallel streams for large databases – note that databases of millions of Persons are routinely processed in this way.
- Process each Person in the SQL result set from the production database:
  - Map the SQL result set to the XML Transaction Document.

- Run the pre-masking Decision Model.
- Auto function for substituting masking data into the Masking Document.
- Run the masking Decision Model.
- Map the XML Transaction Document to SQL for the masked database insertion.

- Once all DMW Runtimes have completed then launch optional dependent job.

Tertiary dependent Job (Optional)

- Execute a SQL script to reapply database Triggers, Constraints, auto numbering and indexes.

Security View – Technical Details

As described, all masking information in the DMW Masking table will be held as an XML fragment within an encrypted database column. Only the original Person identifier is in clear text. At no time within the process is both the original and replacement id available at the same time outside of the DMW code that performs the decryption.

It is the usage of the XML document type of "Masking Data" by a decision model that invokes the automatic DMW retrieval and decryption of masking information as requested within the XML document.

The Advanced Encryption Standard (AES, in CBC+CTS mode with random IV for each file) is used to encrypt the masking data held within the DMW 'Masking' table using an encryption key provided by the customer. More precisely, this 256bit encryption key will be provided by 2 different DMW users, each user providing a full 256bit key which will be bitwise XORed to produce the final key. These 2 users are required to separately sign into DMW under DMW access control to put in their key. This ensures that no single individual knows the encryption key. This encryption key will be stored in an (IDIOM) encrypted form within the DMW database. These two user ids will be stored within DMW and associated with the organization's encryption key.

A propriety encryption key provided by IDIOM and embedded within the code will be used to encrypt the organisations key for storage. It is also used to encrypt the data originally supplied by IDIOM as part of the DMW solution. We recognize that this leaves an exposure to decompiling the delivered IDIOM product code to identify this key, however, this risk is assumed to be acceptable in this context.

This method of Encryption/Decryption is relatively lightweight and does not have a major impact on performance. Only the actual decryption of the XML data column occurs on a per transaction basis.

Initial Provision of the Masking Data

IDIOM provides a table of fake data that will pass standard commercial program validations. This data is preloaded into a static DMW table. This data will be delivered already encrypted using the Idiom proprietary key. Customer specific versions of this data can be inserted if so required.

Creation of the actual 'Masking' table requires the addition of a real Person identifier, and re-encryption of the masking data using the Customer's two part key. This will be a repeatable user requested process

under customer control that will generate the Person identifier's and then assign the fake masking data to each identifier, starting at a random index.

This will ensure that no individual has access at any time to the unencrypted masking information.

Three elements within the masking data have been generated as valid numbers according to their check sum regime. These are ABN, TFN and Bank Account numbers.

Identifying Masked Keys

In certain circumstances it may be necessary to provide a masked test database to vendors for problem resolution. The organisation will need to provide to the vendor the masked key of the record where the problem occurs in the unmasked database.

A special DMW feature is available only to approved staff for this purpose.

The same two people that provided the organisation's encryption key will sign on to the DMW utility function using their DMW user profile. Should either of these users be unavailable another DMW user may be assigned to replace them through an appropriate approval process. These users will be able to enter one at a time the original Person identifier and will be shown the masked id in the test database.

Summary

The data masking capability that is now supported by DMW supports masking of large sets of personal data records, quickly and efficiently across millions of records. Consistency of masking data across foreign keyed relationships is maintained across the full breadth of the database. Using DMW's Windows based parallel processing, it is possible to schedule data masking on any standard Windows machines outside of normal operating hours on very low cost infrastructure.

Data masking that is complete, consistent, large scale, and very cost effective to operate.

DATA MIGRATION

Introduction

Data migration is the process of reading data from one system or version of a system, cleaning, validating, and transforming it, and then writing it back into a new system or version of a system in its new format.

A mix of independent input and output mappers, combined with IDIOM decision models that are found in the IDIOM Decision Manager Workbench [DMW], provide an ideal infrastructure for migration of data between systems or versions of systems.

How it Works

The DMW natively provides the ability to read XML transaction data directly from a file system and write data back to a file system.

If the data is held in relational databases, then the IDIOM Mapper (a component of DMW) provides an easy to configure tool for high speed mapping of relational data to XML, and passing it into the DMW process as per the file system above.

Of course, the Mapper can also map in reverse and insert or update data back into a target system.

Unlike Data Masking, in the case of data migration we use different XML schemas for the input and output documents, with one or more decision models cleaning, validating, and mapping the data between the two schema defined documents.

The IDIOM Decision Manager is an ideal tool for the cleaning and validation process, and it can use the DMW's native 'alerting' capability to notify exceptions, breakages and/or records that could not be migrated as is.

It also has the capability to intelligently transform data between XML documents, including such things as:

- Re-calculate and validate any variables;
- Convert between coding systems (e.g. new look ups to enhanced coding systems, etc);
- Convert between calculation systems (e.g. re-calculate imperial into metric measures, increase or decrease precision, etc);
- Change data types;
- Create, re-collate, re-sequence, and/or redefine data that sits in collections;
- Map multiple fields from a single record into collections and vice versa;
- A multitude of other similar complex tasks.

The result is that data from one system can be converted and re-inserted into a like or unlike system on a large scale. This can be achieved with simple mapping and decision model configurations, and does not usually require development of any bespoke computer code.

Finis

IDIOM Contact Details

Mark Norton | Director | Idiom Limited |
Office +64 9 630 8950 | Mob +64 21 434669 | After Hrs +64 9 817 7165 |
Aust. Free Call 1 800 049 004
1-8 93 Dominion Rd. Mt Eden Auckland 1024 | PO Box 60101 Titirangi Auckland 0642 | New Zealand

Email mark.norton@idiomsoftware.com | Skype Mark.Norton |