

# Profiling Data Best Practices

Case studies and samples of groundbreaking work being done by Axis Technology, LLC in data security

A DATA SECURITY SERIES CASE STUDY

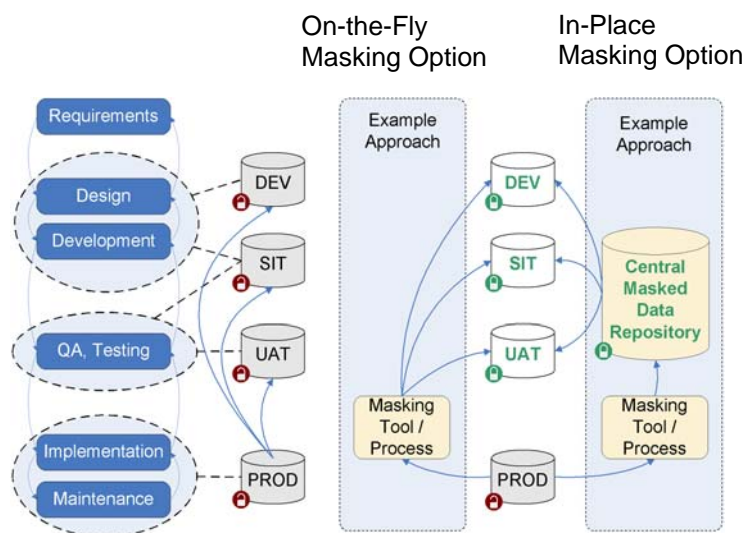
The issue of data privacy can no longer be ignored. The combination of customer outrage around data breaches and new laws and regulations requires companies to have a data privacy policy in place or risk significant fines and negative publicity.

**Introduction** Starting in 2002 with the California data privacy law CA 1386 until recently with the Massachusetts data privacy law MGL 93H, companies are being required to protect customer data and have a data privacy policy in place. Inefficient or non-existent Data Privacy programs result in inconsistent, incomplete, and overlapping use of resources (people, processes, and tools) to meet data privacy objectives. After establishing or updating your data privacy policy, you need to determine the data you ultimately need to mask. This step is called *profiling your data*.

## 1 Data Profiling

The purpose of profiling data is to identify sensitive data elements that require masking by creating an inventory of your data with sensitive data elements (non-public information) identified. You can then review and edit that inventory.

You can mask data on-the-fly or you can provision the data and then mask it (in-place masking). Whichever you choose, you need to profile your data first.



## 2 The DMSuite Profiler

The DMSuite profiler uses two different methods to identify the location of sensitive data:

- **At the metadata level:** searches through the column names in the target database, by querying the database catalog, looking for specific words in column names (for example, column names with "name" in them).
- **At the data level:** looks at the data itself using a sampling algorithm, to see whether there is any sensitive data.

DMSuite then uses that profile information to generate the appropriate jobs that will mask the target database. The user defines the connections to the databases to profile and then uses the DMSuite software to perform the Profiling. When the profiling is complete, the information is stored as profile metadata for DMSuite processing in the locally hosted or network DMSuite database.

In the case of profiling distributed database sources, the source database catalog is read to discover tables and columns that are profiled. After profiling is done, DMSuite can then provision masked tables from the original source. For MVS, copybooks provide the field information.

# Profiling Data Best Practices

A DATA SECURITY SERIES CASE STUDY

## 3 Using Expressions to Profile Data

Expressions let you tell DMSuite how you want to profile data by letting you limit the data to profile based on the criteria you enter in the expressions. For example, you can define an expression that looks for a name or partial name for a column and only profiles data in columns that match that name or partial name.

Description	Expression
Looks for addresses by searching through patterns in the column name	(i:ad(d dress)_line1 ad(d dress)1 city_ad(d dress) ad(d dress)_city address.p[^\^o].*)
Looks for address line information in data	(.*[s]+b(ou)?l(e)?v(ar)?d[s]*.*) (.*[s]+st[.])?(reet)?[s]*.*) (.*[s]+ave[.])?(nue)?[s]*.*) (.*[s]+r(oa)?d[s]*.*) (.*[s]+l(a)?n(e)?[s]*.*) (.*[s]+cir(cle)?[s]*.*)
Looks for address line 2 information in the data	(?i)(.*[s]*ap(ar)?t(ment)?[s]+.*) (.*[s]*s(ui)?te[s]+.*) (c(are)?[s]*[\\\/]?[/]?)o(f)?[s]+.*)

## 4 Profiler Sets

A profiler set is a grouping of expressions for a particular purpose. DMSuite comes with a handful of predefined Profiler sets for HIPAA and PII. In addition, you can define your own custom Profiler sets based on your data.

**Financial Profiler Set**

- First Name
- Last Name
- SSN (Social Security Number)
- Address
- Credit Card Account Number
- Bank Account Number

## 5 Practical Profiling Example

Here's an example of how you might define the data you want to profile.

If you want to look for First Name, define a regular expression to specify how to look for it. If the expression is column-name specific, DMSuite will identify which column names match the pattern specified in the expression.

If DMSuite finds a match, it will tag it as a sensitive column.

Profiling data takes a sample against the column. DMSuite does not look at all rows, but the first n (n being 10,000 rows, 100,000 rows, and so on).

So, if you want to look for First Names across all of your databases, specify the following expression:

```
[Nn][Aa][Mm][Ee]
```

If the expression is at a data level, you can look for common names such as John and Mary:

```
(([Jj][Oo][Hh][Nn])|[Mm][Aa][Rr][Yy]))
```

This expression looks for the names John and Mary in the database.

If DMSuite finds any, it identifies that as a First Name column, which automatically defaults the column to the correct masking algorithm.

You can also search based on format. For instance you can look for a social security number by looking for nine digits of data, with two hyphens (at positions 3,1 and 7,1).