MICRO FOCUS® | white paper

# Productive and Secure Application Testing

## EXECUTIVE SUMMARY

Recent publicity surrounding security breaches in both private enterprises and the public sector has emphasized the need to ensure that all data held by organizations is only used in a legal and compliant manner. While live systems typically already incorporate many types of access controls, the security of access to application test data derived from production data is often far less rigorous. Furthermore, the productivity of the application testing process can be enhanced through an automated and repeatable process to rapidly create a significantly smaller but representative test dataset.

## INTRODUCTION

Enterprise systems share many characteristics, not least the significant volumes of data processed; an amount which is estimated to be growing annually by 125%. At the same time, mission critical applications continue to require change and enhancement in response to business needs. A major factor in the IT rollout of changes to enterprise systems is application testing. Unit, integration, system and user acceptance testing and their associated test data play an important part in the lifecycle of application changes. Still, it is extremely challenging to predictably create artificial test data that can represent all the possible permutations and idiosyncrasies of real data. For this reason, in order to mitigate the risk of disruption to key systems, the majority of major organizations use production-derived data as part of their test process. The use of production-derived data, typically from a backup copy, is understandable and often essential, but can raise challenges and concerns about both the volume of the data itself and the confidential information it represents. This paper seeks to explore those issues in detail and present choices to address these concerns.

> "As concerns around data privacy mount, IT organizations find themselves revisiting their policies around test data. Indeed, customer data and company confidential data straight from production are inappropriate for use in test environments and should be replaced by data that is masked or generated using homegrown or commercial tools"
>
> A Four-Step Program for Compliant Test Data Practices, Forrester

## DATA PRIVACY - CHALLENGES AND CONCERNS

Production systems are protected by an often elaborate and definitive security system, with multiple layers of specified access controls for various features, functions and information contained within the system. The use of production data for testing has the potential to undermine this established security process. Security on test systems cannot be as rigorous. Indeed, engineers will often possess the technical skills to access and view the test data, directly bypassing all security protocols. Programmers are often required to debug application problems and investigate system behaviour as part of their day to day tasks. This means that if unaltered production data is used as part of the testing process, it would inevitably be visible to multiple application development staff.

## BUSINESS CONCERNS WITH DATA PRIVACY COMPLIANCE

> "Over 262 million personal information records have been compromised in breaches of data security reported to the authorities in the USA during the last four years."
>
> A Chronology of Data Breaches, Privacy Rights Clearinghouse, 2009

The consequences of failing to secure customer data are evident on a regular basis, not only in the trade press, but on national and international news bulletins. Ensuring the privacy of clients' data is a key requirement for most organizations, and the cost of failing to protect sensitive customer data can be very high. The use of unaltered production data in the testing process presents a number of risks:

- Test environments are not physically as secure as production systems. Sensitive data could for example be downloaded to laptops which are subsequently stolen
- Employees could misuse sensitive information such as credit card details and engage in fraudulent activity
- The drive to outsource and offshore application development and testing work means that organizations will have little control over who gets to see their the test data
- Enterprise systems share many characteristics, not least the significant volumes of data processed; an amount which is estimated to be growing annually by 125%.
- Adverse publicity concerning a breach of privacy can severely impact customer confidence in an organization which can translate into significant loss of business and a fall in stock price

> "The average cost per security breach for each reporting company was $6.7 million, and the average cost per lost customer record was $202, compared to $182 two years ago"
>
> The Cost of Data Breach, Ponemon Institute, 2008

| Law / Standard | Territory / Vertical | Notes |
|---|---|---|
| Health Insurance Portability and Accountability Act (HIPAA) | USA / Healthcare | Detailed legislation which includes names, geographic elements, telephone numbers, social security numbers and health plan numbers |
| Federal Information Security Management Act (FISMA) | USA / Federal | Provides sets of controls for information security programs, and other essential requirements and guidance |
| Personal Information Protection and Electronic Documents Act | Canada | Local legislation |
| Privacy Act | New Zealand | Local legislation since 1993 |
| BASEL II | Banking | Recommendations on banking laws and regulations |
| New York State | USA | Local legislation that came into effect January 1st 2008 - social security numbers cannot be used in test databases |
| Data Protection Act | UK | Wide measure on protection and usage of personal data |
| ISO/IEC 27002 | | Security techniques code of practice for information security management |
| European Union Data Pro-tection Directive 95/46/ED | EU | Regulatory framework for different national data protection laws in the EU |
| Italian law 675/1996 | Italy | Defines personal data as all the identifying information concerning the individual, legal entity, institution or association |
| Sarbanes-Oxley | USA | US Federal regulation governing corporate and accounting standards |
| Privacy Amendment Act | Australia | Local legislation since 2000 |
| Markets in Financial Instruments Directive | EU / Finance | EU law providing a harmonised regulatory framework for investment services |
| Payment Card Industry (PCI) | Banking / Retail | Requires protection for key information such as account numbers, cardholders names, service codes and expiration dates |

## LEGISLATION GOVERNING DATA PRIVACY

The area around data privacy is covered by a range of legislation and compliance criteria on a national, international and industry basis. In some cases a compliance breach of national legislation could render a corporation subject to significant fines, and potentially directors or other individuals subject to imprisonment. Meanwhile, major industry standards such as Basel II in Europe and HIPAA in the US are effectively barriers to operation in major areas of business.

Thus organizations have no choice but to navigate the myriad of standards and laws that could apply to their business operations. The depth and specifics of these regulations varies greatly, but they will typically specify a range of criterion for the 'Deidentification' or 'Desensitization' of information. In particular, information that must be removed, obscured, or removed from its context includes personal identity information (names, dates of birth, addresses, social security numbers, health/insurance numbers, photographs, etc.) and financial information (credit card numbers, expiration dates, bank account numbers, etc).

## OUTSOURCING AND DATA PRIVACY

Increasingly, development, maintenance and testing of enterprise applications is not being carried out by an organization's own employees. Instead, this work is outsourced, typically overseas, to countries such as India and China. Outsourcing raises a number of additional factors, some more obvious than others. Quite clearly, outsourcing means that additional legislation may apply. For example the consequences of breaching US Safe Harbor directives can be severe.

Other concerns are potentially more subtle but equally as significant. It will often be the case that the outsourced developers will have less specific business knowledge or understanding about the execution of the application. While working on individual elements of an application, the overall performance of a that application could be degraded and problems could be introduced that are not detectable by basic unit testing.

In many situations due to legal concerns, no more than simple test data is passed to outsourcers. As a result, the code that is returned back to the organization may not be of a high or proven quality. When full testing is carried out, numerous additional problems or errors are revealed that can cause further delay in the rollout of the new functionality.

## DATA MASKING OVERVIEW

The process of data masking is designed to "de-identify" or "de-sensitize" data, such that the data remains based on real information but no longer has any practical usage or application. In other words, it is now data rather than information. This involves the obfuscation of any information subject to any internal procedure, national law or industry regulatory compliance, such that if this data is subsequently used in an inherently insecure context such as application testing, the data no longer contains meaningful private or confidential information in any way.

The number of fields that will require masking will of course vary greatly by application, organization and legislative requirements. In general it would be preferable only to mask columns that require masking to maintain a straightforward process that ensures the integrity of the data. Furthermore there is no single answer to the correct masking methodology or algorithm; indeed it could well be argued that a preferable approach is not to employ a single technique in masking data.

Masking in a large enterprise may well include elements unique to that organization and combine multiple techniques, but as building blocks common masking techniques include:

### › Simple masking

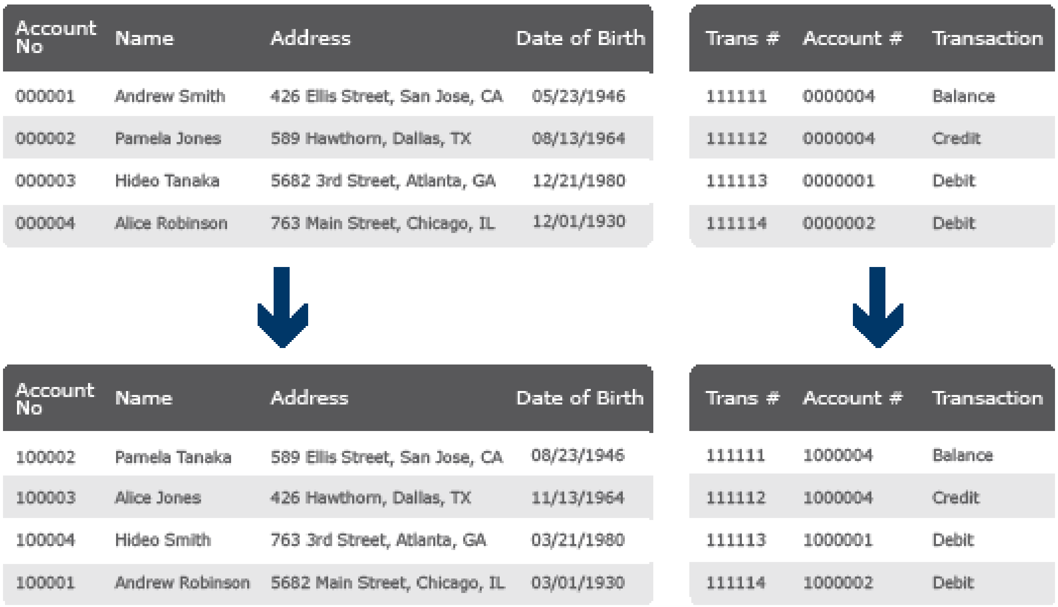
In essence sensitive data is simply replaced with a static set of null values such as XXXX or 9999. This technique is sometimes used in simpler manual masking processes, and it is secure in that the original information is obviously effectively masked. However this approach departs from production-like data and increases the risk both of testing problems resulting from this data and, more importantly, of testing not detecting problems that will occur when the applications are executed against live data.

### › Numeric manipulation

At its simplest this approach basically increments or decrements the data by a given range. For example an order value could be increased by 5% or the data could be aged by adding say 1000 days to a date of birth. The simpler versions of this approach should however be treated with caution; a simple algorithm could be deciphered and again the data may no longer represent the production data characteristics. More complex numeric manipulation based on a set of circumstances can however overcome some of these limitations.

### Data substitution

This is a commonly used approach, and if used well can be extremely effective. Data is substituted with an alternative which can be determined randomly or through more sophisticated replacement mechanisms. The integrity of this approach is dependent on the data substituted; key of course is to preserve the usefulness of the data while obscuring its information value. In a form of data shuffling the data can be exchanged from different rows in the database, so one account number for example is now associated with the name of another account holder, and so on. Data substitution has the distinct advantage of preserving the variations and idiosyncrasies of the original production data. This approach can be enhanced by using external data sources for substitution, for example a list of names from a phone book in place of customer names, other valid zip or postal codes for a given town, etc. In some cases the application will itself validate the integrity of a piece of data such as a postal address, making the need for intelligent substitution critical.



| Account No | Name | Address | Date of Birth |
|---|---|---|---|
| 000001 | Andrew Smith | 426 Ellis Street, San Jose, CA | 05/23/1946 |
| 000002 | Pamela Jones | 589 Hawthorn, Dallas, TX | 08/13/1964 |
| 000003 | Hideo Tanaka | 5682 3rd Street, Atlanta, GA | 12/21/1980 |
| 000004 | Alice Robinson | 763 Main Street, Chicago, IL | 12/01/1930 |

| Trans # | Account # | Transaction |
|---|---|---|
| 111111 | 0000004 | Balance |
| 111112 | 0000004 | Credit |
| 111113 | 0000001 | Debit |
| 111114 | 0000002 | Debit |

| Account No | Name | Address | Date of Birth |
|---|---|---|---|
| 100002 | Pamela Tanaka | 589 Ellis Street, San Jose, CA | 08/23/1946 |
| 100003 | Alice Jones | 426 Hawthorn, Dallas, TX | 11/13/1964 |
| 100004 | Hideo Smith | 763 3rd Street, Atlanta, GA | 03/21/1980 |
| 100001 | Andrew Robinson | 5682 Main Street, Chicago, IL | 03/01/1930 |

| Trans # | Account # | Transaction |
|---|---|---|
| 111111 | 1000004 | Balance |
| 111112 | 1000004 | Credit |
| 111113 | 1000001 | Debit |
| 111114 | 1000002 | Debit |

*Simple data masking example, utilizing simple substitution and some basic numeric manipulation, preserving referential integrity*

Whatever techniques are employed, it is critical to define an appropriate action for each sensitive data element and to be able to repeatedly apply a consistent masking process which is propagated through all related data sources.

## BEYOND PRIVACY – DOES IMPROVING TEST DATA REALLY MAKE A DIFFERENCE?

Regardless of whether traditional waterfall or increasingly influential agile development methodologies are employed, the role played by system, integration and user acceptance testing – a comprehensive test of all application elements integrated and deployed against a realistic test data set – remains critical in enterprise IT deployments. Often at these stages, problems that occur at the end of a project, while certainly preferable to production failures, can be costly and add risk to business deadlines.

Delivering appropriate and timely test data is a key pre-requisite in this process. Considerations include:

### Availability of test data

While the speed of ‘data aging’ will vary according to business usage, it is essential to ensure that current test data is used during the testing process. If a regular refresh cycle is not in place and resourced, the creation of new test data from the latest production data will be a key factor to plan and schedule.

### Size of test data

In many ways a complete copy of production data is most desirable, but the implications of processing for example 5 terabytes, 15 terabytes or even 50 terabytes of data deserves careful consideration. The greater the size of data, the longer a system test will take - a difference that will be multiplied if a retest is necessary due to a test failure. Beyond the obvious amount of time for tests themselves to execute, the sheer handling and replication of such huge data stores is time consuming and unwieldy. As a result, most organizations implement a degree of ‘subsetting’ to reduce the data size.

- **Quality of test data**

  It perhaps seems obvious – test everything and you will optimize quality. In reality, considerations are more complex. A large and unmanageable test data set may result in compromises in its execution and jeopardize the ability to adequately test, and as discussed earlier, may not be acceptable for privacy reasons. However, as soon as the data is subsetted or masked, the issue of quality arises. Specifically, does the reduced and secured test data remain a sufficiently complete and integral representative data set in order to perform system testing? Is the referential integrity of the data intact? Less evident perhaps is predictability - if the reliability of test data varies each time it is processed, this has the potential to severely compromise application testing.

Reviewing the effectiveness of these core points can substantially improve the efficiency of key stages of the application testing process. As practices become optimal, this improved operation has the potential to deliver real business benefits, such as:

- **Productivity gains delivering more rapid time to market**

  The timely and predictable availability of appropriate test data mitigates a key task in application change rollout. If the size of that test data set is substantially reduced (by between 50% and 90%) then the time taken to perform application testing can be cut, further reducing this key project phase. The more rapid and repeatable this process becomes, the more application test cycles can be optimized, increasing the productivity of individuals and the process overall, thereby reducing the time required to deliver new business changes.

- **Improved quality resulting in less system failures and problems**
  No-one wants to find bugs in the late stages of testing. However, the earlier and more accessible appropriate test data is available, the earlier problems will be detected if they exist. It remains vastly preferable to find a problem in system testing than in production. Towards the end of the testing cycle the right test data can be argued to be the single most important factor in testing to determine 'production-ready' quality. Certainly rigorous code coverage, change management, QA procedures and diligent engineers cannot fully replace this core requirement.

- **Reduced usage of MIMIPS and storage delivering internal cost savings**
  Processor time costs money. This may be explicitly cross-charged within an organization or it could be less visible, but few enterprises consistently have excess capacity to be used unproductively. It is not atypical for up to 50% of mainframe MIPS to be consumed for development purposes rather than production business use. Indeed, executing a large system test with, say, 30 terabytes of data remains an expensive operation, even considering the gradual consolidation of processing costs. Additionally, if the data store is very large, then multiple replications of test data will add to storage costs. The ability to substantially reduce (by up to 90%) the test data set used can notably reduce MIPS and storage used in the application testing process, delivering internal and actual cost savings.

## BUILD YOUR OWN – ADVANTAGES AND PITFALLS

Given that securing test data is often obligatory, and sub-setting that same data is often critical to ensure effective testing, the question arises: by what means can this be most effectively delivered? Traditionally when this need has arisen in many organizations, an in-house solution has been crafted. In this sense an internal solution can deliver advantages:

- Rapidly addresses an immediate, simple requirement for a one-time or occasional operation
- Offers a tailored approach to current needs and disparate data stores
- Leverages internal data knowledge

The effectiveness of such a solution will naturally be dependent on both the requirement and the level of resource invested into creating a tailored infrastructure around it. Common pitfalls include:

- Resourcing - most internal processes are semi-manual in their operation. They include a batch and potentially programmatic element but require significant resources to prepare and execute the process. This can impact the ability of an organization to meet requirements for timely masked and reduced test data

- Complexity - the processes of de-identification and sub-setting data are not straightforward. It is necessary to build an understanding of the data, the interrelationship between data elements across many hundreds of database tables both on and off the mainframe, and of VSAM files. An incomplete understanding of these relationships, and the lack of a strategic repository to represent this knowledge, will typically lead to incomplete masking and minimal or even erroneous sub-setting. Intensive resourcing will improve the situation but is unlikely to achieve an optimal position.

- Repeatability - a risk with manual intervention within a process is that the operation of that process may not be well defined or consistent on repeated operations. This adds risk to a project, and could result in inconsistent or ineffective test data which itself would jeopardize the testing process and schedule.
- Effectiveness - without any infrastructure to classify and inventory data, and without considerable effort in leveraging information during de-identification and sub-setting, there is a substantial risk that masking may not be complete, and the data size reduction may only be a fraction of that potentially achievable.
- Maintenance and extensibility - strategic data stores are often changed and extended. Depending on the changes, the process for de-identification and masking may require either maintenance or potentially extension. Such internal solutions can be difficult to rapidly and reliably extend, especially if the engineers who originally crafted the solution are unavailable. Building your own solution may well have been the right historical decision for many organizations; it provided a rapid, albeit tactical, response to an immediate need. However, unless the organization is very small, it has very simple requirements, or can commit very substantial resources, it is unlikely that an internal solution will represent a long-term strategy to deliver compliance and productivity in application testing.

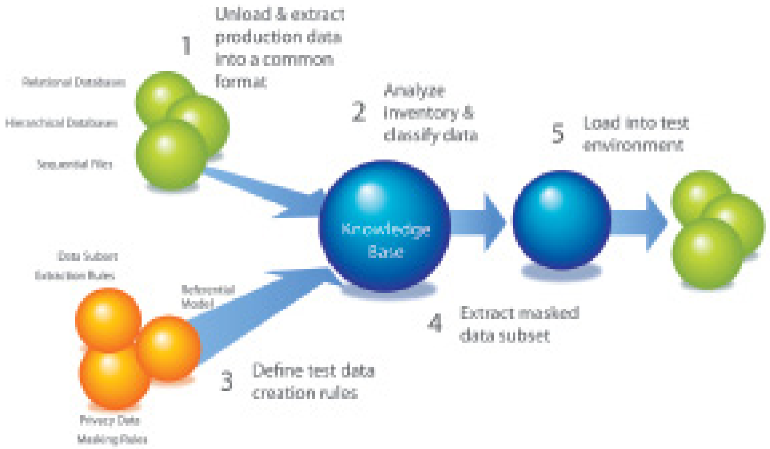
## MICRO FOCUS DATA EXPRESS

Micro Focus Data Express allows organizations to improve the testing process in a secure and compliant environment by enabling the creation of de-identified and significantly reduced test data in an automated and repeatable infrastructure. This facilitates the availability of a regular "heartbeat" of predictable, secure and manageable data to application testing groups, ensuring test data is available as and when required. By regularly providing this robust, secure and productive test data environment, Data Express enables enterprises to use a single tool to both meet stringent privacy and compliance needs while rapidly creating a test data environment often less than 10% of the production data size. To achieve this, a five-step process can be defined:

- Build the knowledge base from the production data (for example a backup copy) through direct access or an unload process
- Analyze, inventory and classify the data in the knowledge base. This provides the key information for subsequent steps
- Define the extraction patterns and rules delivering repeated extraction schemes. These rules are reused in all subsequent extractions
- Perform the actual data extraction. This is a single process for both masked and subsetted data
- Load the reduced and secured test data into the test environment for testing against normal procedures

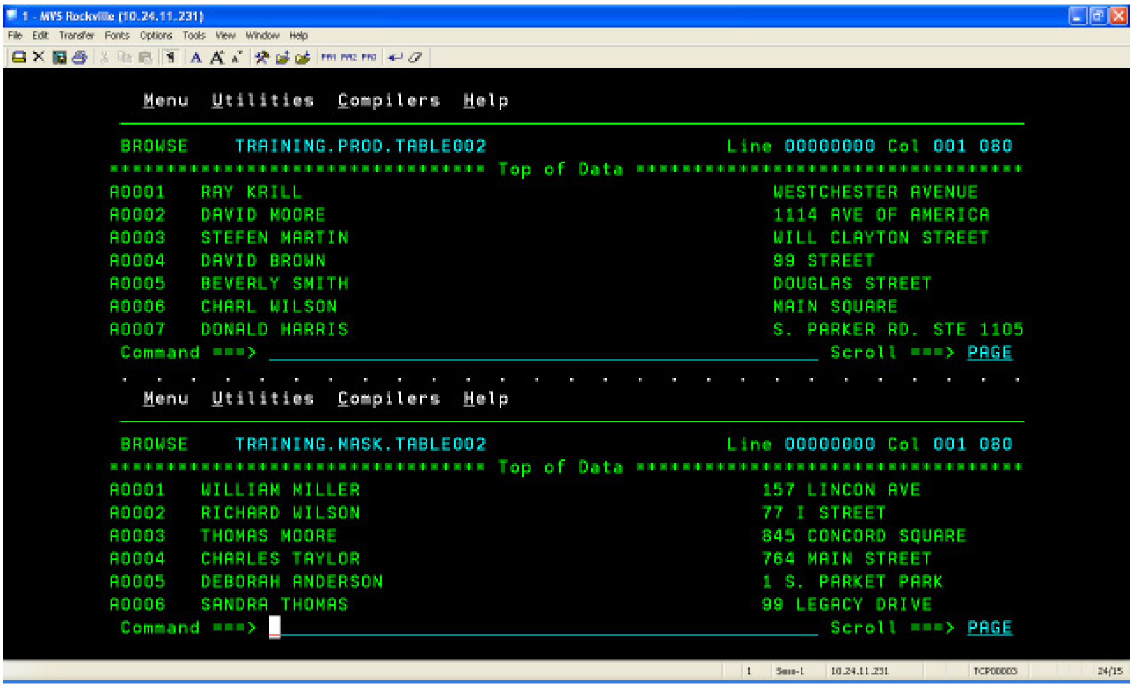
## A STRATEGIC REPOSITORY – SECURE, ACCESSIBLE, INTEGRATED AND EXTENSIBLE

The Builder component of Micro Focus Data Express allows an inventory of organizational data to be taken, collated and centrally stored in a knowledge base. Strategic data structures including IMS DB, DB2, VSAM, ADABAS, Oracle and SQL Server are collated in this central repository, accessible from a single administrator desktop. The knowledge base doesn't store the customer data itself, but rather the structure and information about the data. The information within the knowledge base is a key pre-requisite to rigorous and consistent sub-setting and masking across disparate data sources, and provides a strategic and secure repository. Specifically the process of understanding the nature, contents and relationships of enterprise data is critical to successful data reduction and de-identification.


1
Unload & extract production data into a common format
Relational Databases
Hierarchical Databases
Sequential Files
2
Analyze Inventory & classify data
5
Load into test environment
Knowledge Base
Data Subset Extraction Rules
Referential Model
4
Extract masked data subset
3
Define test data creation rules
Privacy Data Masking Rules

Data Express includes:

- Rapid inventory, classification and analysis of data
- Powerful sampling and 'fingerprinting' capability which enables the understanding and classification of data fields, for example as dates of birth, names, etc.
- Unique classing and super-classing features of Data Express support the creation of common actions for groups of fields with homogeneous characteristics, accelerating the delivery of consistent rules and actions for data changes
- Common extraction of masked and reduced data


1 - MVS Rockville (10.24.11.231)
File Edit Transfer Fonts Options Tools View Window Help
1
Sess-1
10.24.11.231
TCP00003
24/15

```
    Menu  Utilities  Compilers  Help
 ------------------------------------------------------------------------------
 BROWSE    TRAINING.PROD.TABLE002                      Line 00000000 Col 001 080
********************************* Top of Data **********************************
A0001   RAY KRILL                                        WESTCHESTER AVENUE
A0002   DAVID MOORE                                      1114 AVE OF AMERICA
A0003   STEFEN MARTIN                                    WILL CLAYTON STREET
A0004   DAVID BROWN                                      99 STREET
A0005   BEVERLY SMITH                                    DOUGLAS STREET
A0006   CHARL WILSON                                     MAIN SQUARE
A0007   DONALD HARRIS                                    S. PARKER RD. STE 1105
 Command ===> ________________________________________________ Scroll ===> PAGE
 .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
    Menu  Utilities  Compilers  Help
 ------------------------------------------------------------------------------
 BROWSE    TRAINING.MASK.TABLE002                      Line 00000000 Col 001 080
********************************* Top of Data **********************************
A0001   WILLIAM MILLER                                  157 LINCON AVE
A0002   RICHARD WILSON                                  77 I STREET
A0003   THOMAS MOORE                                    845 CONCORD SQUARE
A0004   CHARLES TAYLOR                                  764 MAIN STREET
A0005   DEBORAH ANDERSON                                1 S. PARKET PARK
A0006   SANDRA THOMAS                                   99 LEGACY DRIVE
 Command ===> ________________________________________________ Scroll ===> PAGE
```

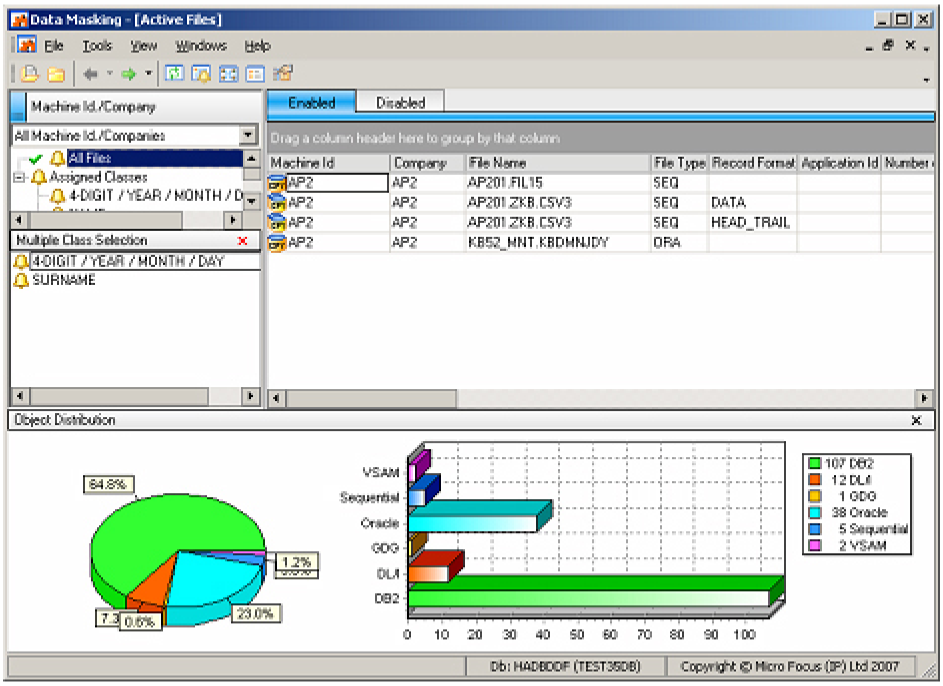
*Before and after de-identified data*

## EFFECTIVE DATA MASKING FOR PRIVACY AND COMPLIANCE

We've discussed earlier the critical requirement for data de-identification or masking in order to comply with international privacy laws, regulatory standards and common best practice. Data Express supports the obfuscation and masking of sensitive and private data elements to build an anonymous and compliant test data environment.

Data Express masking capability includes:

- Classification, cataloguing, inventory and masking of personal and sensitive data
- Predefined masking routines for common fields such as name, surname, ID codes, email address, telephone number and postal address
- Customized masking routines to build user defined functions for unique rules and masking
- Compliance with standards such as HIPAA, BASEL II, PCI, Sarbanes-Oxley and many others
- Data Class feature providing common actions for related fields



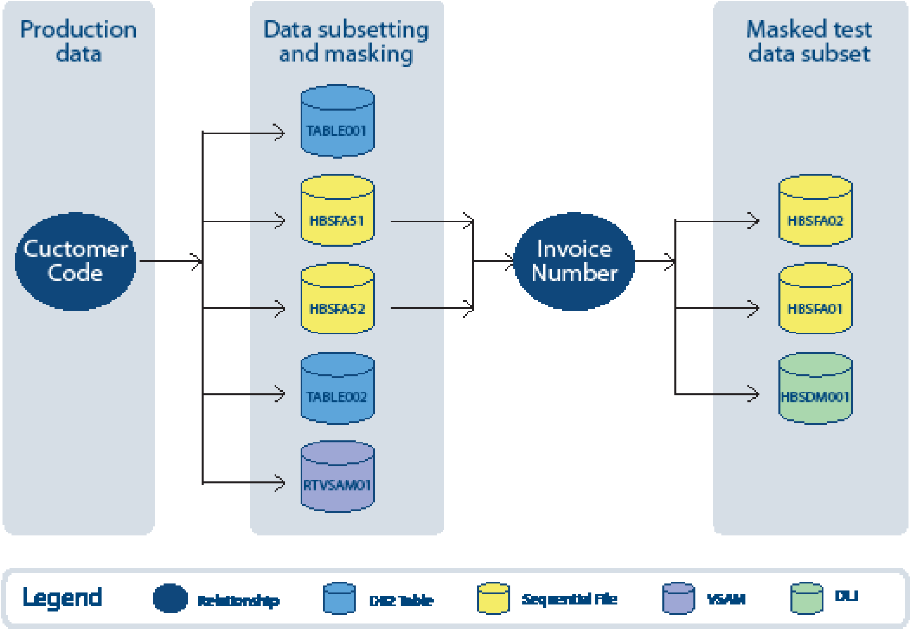
| Machine Id | Company | File Name | File Type | Record Format | Application Id | Number |
|---|---|---|---|---|---|---|
| AP2 | AP2 | AP201.FIL15 | SEQ | | | |
| AP2 | AP2 | AP201.ZKB.CSV3 | SEQ | DATA | | |
| AP2 | AP2 | AP201.ZKB.CSV3 | SEQ | HEAD_TRAIL | | |
| AP2 | AP2 | KB52_MNT.KBDMNJDY | ORA | | | |

*Data masking – multiple class selection*

## Reduce test data by 90%

Data Express allows organizations to extract a consistent and congruent subset of data to reduce, often substantially, volumes of data in the test environment. Data Express analyzes large databases to derive the right amount of data for testing in the best possible time, generating reduction rules from a data model file created earlier by the Data Modeller.

The Data Express knowledge base supports the understanding and definition of cross-relationships within the data. It ensures referential integrity is maintained by propagating changes consistently across all related databases and files. By utilizing these congruent relationships between data of different sources to apply repeatable and not random mechanisms, consistency and logical data integrity are ensured, and unique constraints of simple and composite keys are preserved. An extraction rules wizard supports the definition of complex extraction patterns as part of the creation of an automated and repeatable sub-setting process. These rules are applied during the extraction process to build the resulting test data.



*Defining the relationship between disparate data sources*

## SUMMARY

Most enterprises with non-trivial application portfolios will have both a legal and a cost requirement to investigate and employ an appropriate solution for management of test data. Micro Focus Data Express provides an effective solution for the creation of secure and substantially reduced application test data. It enables an organization to ensure both legal and regulatory compliance, and optimize time to market of new business requirements by improving the productivity of the testing process. The same reduced and de-identified data can also be used more widely for training and other non-production purposes.

### About Micro Focus

Micro Focus, a member of the FTSE 250, provides innovative software that allows companies to dramatically improve the business value of their enterprise applications. Micro Focus Enterprise Application Modernization, Testing and Management software enables customers' business applications to respond rapidly to market changes and embrace modern architectures with reduced cost and risk.

**For additional information please visit: www.microfocus.com**