Testen mit Produktionsdaten – Fluch oder Segen?

Thomas Baumann

Die Mobiliar

ISACA After Hours Seminar
Dienstag, 30.Oktober 2012
Agenda

• PART I: Copy from Production to Test
  • Requirements
  • Solutions and experiences @Mobiliar
• PART II: The Data Masking Process
  • Requirements
  • Solutions and experiences @Mobiliar
• PART III: The Pre-Production Environment
  • Examples of pre-prod usage @Mobiliar
Swiss Mobiliar
(major Swiss Insurer for all sectors)

- HQ located in Bern, Switzerland
- 4,500 employees
- IT:
  - 200+ application developers
  - DB2 zOS as strategic DBMS for OLTP
  - DWH and CRM (Siebel) running on Oracle

Some typical Swiss attributes

- high precision
- excellence in winter sports
Disclaimer

• The Information contained in this presentation has not been submitted to any formal Swiss Mobiliar or other review and is distributed on an ‘as is‘ basis without any warranty either expressed or implied. The use of this information is the user’s responsibility.

• The procedures, results and measurements presented in this paper were run in either the test and development environment or in the production environment at Swiss Mobiliar in Bern, Switzerland. There is no guarantee that the same or similar results will be obtained elsewhere. Users attempting to adapt these procedures and data to their own environments do so at their own risk. All procedures presented have been designed and developed for educational purposes only.
Why Copying from Prod to Pre-Prod

- All data is valid and consistent
- All production conditions are in pre-prod
- Pre-Prod data size corresponds to production
- Pre-Prod data masked (see Part II)
- Alternate Solution if Pre-Prod data not masked:
  - Pre-Prod Security Identical to Production Security
    - Compression/Encryption
    - Firewalls
    - Breach Detection/Notification
    - Security Awareness / Training
    - Access limited to strict minimum of people
Daten kopieren: Testfragen (1 richtig, 3 falsch)

a) Testdaten unterliegen keinerlei gesetzlichen Anforderungen bezüglich Datenschutz, sofern sie klar als Testdaten gekennzeichnet sind.

b) Der Export von Testdaten in Nachbarländer ist ohne Einschränkungen erlaubt.

c) Es gibt keinen gesetzlichen Unterschied zwischen Testdaten und Produktionsdaten.

d) Testdaten sind implizit stets für internen Gebrauch klassifiziert, nie aber als vertraulich.
Think before you Copy

- 46+ US jurisdictions with laws about protecting financial data
  - 1996: HIPAA (Health Insurance Portability and Accountability Act)
  - 1999: Gramm-Leach Bliley Act
  - Payment Card Industry Data Security Standard (PCI DSS)
- EU Data Processing Directive
- Company Reputation and Consumer Trust
- Legally no difference between production and test data
Security Breach Hall of Shame

- [www.privacyrights.org](http://www.privacyrights.org)
- Chronology of Data Breaches

<table>
<thead>
<tr>
<th>Date Made Public</th>
<th>Name</th>
<th>Entity</th>
<th>Type</th>
<th>Total Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 11, 2012</td>
<td>Northwestern Mutual and One America - American United Life, Indianapolis, Indiana</td>
<td>BSF</td>
<td>INSD</td>
<td>3,000</td>
</tr>
<tr>
<td>August 23, 2012</td>
<td>Charter One, Dollar Bank, Fifth Third, First Mort, Key, PNC, Total Merchant Services, Cleveland, Ohio</td>
<td>BSF</td>
<td>INSD</td>
<td>Unknown</td>
</tr>
<tr>
<td>August 23, 2012</td>
<td>State Farm Insurance, Bloomington, Illinois</td>
<td>BSF</td>
<td>INSD</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

A former financial planner stole sensitive information from approximately 3,000 clients and used it to open new accounts, make purchases, receive cash advances, and reroute client mail until his arrest in August of 2011. Client names, Social Security numbers, contact information, and financial account information were exposed. He was sentenced to two years in prison and three years of probation. He will also have to pay $48,488.66 in restitution.

Information Source:
- Databreaches.net

PART I: Copy from Production to Test
Copy From Production to Test: Scope

PART I: Copy from Production to Test
Copy From Production to Test: Scope (cont.)

PART I: Copy from Production to Test
Data Replication (Scope: Customer Data)
Other Requirements

- As close to no-interruption of production as possible
- Only Part of DB instance (selected tables)
- Efficient processing (source + target)
- Minimum manual intervention
- Restartable
- Support of different schema names in source and target
- Support of different schema structures
  - OK if done before and/or after the copy process
- Repeatable on single-object level
- Data masking (see second part of presentation) supported
Some Solutions: The Big Picture

Unload / Load

Cross-Load

Clone Saves

PIT Recover

Cloning Solutions

out of the box

additional licenses

PART I: Copy from Production to Test
The Copy Process: Summary

- Many solutions available
- Easy to use
- Not very expensive software
- Capture & Replay solutions to repeat (production) workload also available
Agenda

✓ PART I: Copy from Production to Test
  ✓ Requirements
  ✓ Solutions and experiences @Mobiliar

• PART II: The Data Masking Process
  • Requirements
  • Solutions and experiences @Mobiliar

• PART III: The Pre-Production Environment
  • Examples of pre-prod usage @Mobiliar
Data Masking Requirements

- Legal privacy requirements define scope
  - @Mobiliar DB2 core data: Different columns of 19 tables identified
  - Irreversible scramble (not just “encryption”)
  - No way for the original production data to be recreated

- Performance requirements of pre-production
  - No loss of data correlation within tables
  - No loss of data correlation among tables
  - Column data distribution identical to production environment

Can't turn the crank backwards and get the pig back
Row-oriented vs. Column-oriented Masking

Row-Oriented Masking

<table>
<thead>
<tr>
<th>Emp_no</th>
<th>Dept_id</th>
<th>Hire_date</th>
<th>Emp_in</th>
<th>Emp_fn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2001-01-01</td>
<td>Smith</td>
<td>Bob</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2002-02-01</td>
<td>Jones</td>
<td>Jim</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2002-05-01</td>
<td>Young</td>
<td>Sue</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2003-02-01</td>
<td>Stamle</td>
<td>Bill</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1999-06-15</td>
<td>Aurora</td>
<td>Jack</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>2000-08-15</td>
<td>Jung</td>
<td>Laura</td>
</tr>
</tbody>
</table>

Column-Oriented Masking

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2001-01-01</td>
<td>2002-02-01</td>
<td>2002-02-01</td>
<td>2002-02-01</td>
<td>2002-02-01</td>
</tr>
</tbody>
</table>

PART II: The Data Masking Process
Row-Oriented vs. Column-Oriented Masking

• **Row-Oriented Masking**
  - Scope of masking algorithm is one single row
  - Difficult to preserve data correlations among many rows while maintaining high data privacy

• **Column-Oriented Masking**
  - Scope of masking algorithm are all rows’ values of a single table column, or of a group of columns of the same table if dependent from each other
  - Data distributions can be maintained
  - If necessary, additional row-masking possible (social security number, financial record/credit/debit card information)
Daten maskieren: Testfragen (1 richtig, 3 falsch)

a) Zeilen- (row-) basierte Anonymisierung und Spalten- (column-) basierte Anonymisierung schliessen sich gegenseitig aus.

b) Es genügt, Daten spaltenbasiert zu anonymisieren, eine zusätzliche Anonymisierung auf Zeilenebene bringt keine erhöhte Sicherheit.

c) Die Anonymisierung auf virtueller Ebene (Views, data masking) genügt, falls stets via Views zugegriffen wird und keine Benutzerrechte direkt auf Tabellen Ebene existieren.

d) Die Anonymisierung auf virtueller Ebene genügt nicht, da Sicherungen auf der physischen Ebene (tablespaces in der Datenbanksprache) durchgeführt werden, und bei bloss virtueller Anonymisierung diese in ihrer ursprünglichen unverschlüsselten Form sichtbar bleiben.
### Column-Oriented Masking: The Basic Idea

<table>
<thead>
<tr>
<th>CustNo</th>
<th>Name</th>
<th>FirstName</th>
<th>Phone</th>
<th>Year of Birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>Baumann</td>
<td>Thomas</td>
<td>+4179...34</td>
<td>1963</td>
</tr>
<tr>
<td>456</td>
<td>Hrle</td>
<td>Namik</td>
<td>+49.........</td>
<td>1958</td>
</tr>
<tr>
<td>789</td>
<td></td>
<td></td>
<td>+33.........</td>
<td></td>
</tr>
<tr>
<td>012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>345</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CustNo</th>
<th>Name</th>
<th>FirstName</th>
<th>Phone</th>
<th>Year of Birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td></td>
<td></td>
<td>+33</td>
<td>1963</td>
</tr>
<tr>
<td>456</td>
<td></td>
<td></td>
<td>+1..........</td>
<td>1958</td>
</tr>
<tr>
<td>789</td>
<td>Hrle</td>
<td>Namik</td>
<td></td>
<td></td>
</tr>
<tr>
<td>012</td>
<td>Baumann</td>
<td>Thomas</td>
<td>+49.........</td>
<td></td>
</tr>
<tr>
<td>345</td>
<td></td>
<td></td>
<td>+4179...34</td>
<td></td>
</tr>
</tbody>
</table>
Stream of Scripts for row-based data masking

- Complex process, based on application-know how
  - Takes 5-10 times longer than simple copy from prod to pre-prod
The Data Masking Process: Summary

- Needs application Know How
- Row-based masking takes 5-10x longer than the copy process
- Automated Solutions available for row-based processing
- Solutions available for data masking on virtual layer
Agenda

- PART I: Copy from Production to Test
  - Requirements
  - Solutions and experiences @Mobiliar
- PART II: The Data Masking Process
  - Requirements
  - Solutions and experiences @Mobiliar
- PART III: The Pre-Production Environment
  - Examples of pre-prod usage @Mobiliar
The Pre-Production Environment

- Continuous Query Performance Monitoring
  - Goal: Compare, predict, optimize access paths at pre-prod
- What-If Performance Analysis
  - Goal: Avoid inefficient access path (often caused by correlation between tables not reflected in the statistics data)
- Performance Prediction
- IT Performance (Response Times) is one of the three CIO’s top priorities
Daten nutzen: Testfragen (1 richtig, 3 falsch)

a) Auf einem anonymisierten Testdatenbestand sind keine Benutzer Zugriffs Einschränkungen mehr notwendig.

b) Auf einem anonymisierten Testdatenbestand sind keine Benutzer Zugriffs Einschränkungen mehr erlaubt.

c) Auf synthetisch generierten Testdaten sind keine Zugriffseinschränkungen notwendig.

d) Der Zugriff auch auf anonymisierte Testdaten sollte auf einem rollenbasierten Zugriffskonzept basieren.
Summary

- Copy from Production to Pre-Prod Test
  - Efficient and non-disruptive solutions available
- Data Masking
  - Need to understand how applications process data
- Pre-Production Environment Benefits
  - What-If Analysis for Performance Issues
    - rat (=real application testing)
  - Enables Production Performance Prediction
Thomas Baumann
Die Mobiliar
thomas.baumann@mobi.ch
Testen mit Produktionsdaten – Fluch oder Segen?