Datenanonymisierung und darüber hinaus

Miroslav Jakovljevic
Libelle AG
Stuttgart

Schlüsselworte
Datenanonymisierung. Datenmaskierung. Oracle & SAP.

Einleitung
Echtdaten zu anonymisieren und dabei realistische Daten beizubehalten, ist eine Herausforderung, sowohl für jeden betroffenen Fachbereich, als auch für die IT eines Unternehmens. Die wichtigsten Einsatzgebiete hierfür sind der Aufbau und Betrieb von Entwicklungs-, Test- und Schulungssystemen.
Besonders bei Tests stehen Unternehmen vor einer kritischen Situation: Auf der einen Seite muss das Testsystem realistische Daten enthalten, um gültige Tests zu ermöglichen, auf der anderen Seite dürfen keine sensiblen Produktivdaten im Testsystem sein.

Also was tun? Personenbezogene Daten für Tests einfach im Original benutzen?
Aus Sicht des Datenschutzes und interner Funktionen (z.B. Personal, Recht, Compliance Management,…) ist dieses Vorgehen keine gute Idee. Zumal Gesetzgeber, Aufsichtsbehörden und andere Institutionen immer mehr und immer schärfere Regeln mit empfindlicheren Strafen erlassen. Doch nicht nur das, es ist auch ein Vertrauensbruch gegenüber Kunden, Geschäftspartnern und Mitarbeitern.

Abhilfe verschafft die Anonymisierung von Daten. Sie ist eine hervorragende Methode, Testdaten gemäß den geltenden Anforderungen bereitzustellen ohne Risiken im Hinblick auf Datenschutz, Sicherheit und Compliance.
Das klingt erstmal einfacher, als es ist.

Wir zeigen die vielfältigen möglichen Gründe für Datenanonymisierung auf und klären welche Stolpersteine es zu bewältigen gilt.
Wir beleuchten, welche Punkte Unternehmen bei der Auswahl der Lösung berücksichtigen sollten und zeigen Lösungsmöglichkeiten auf.

Anforderungen
Zuerst müssen die Anforderungen an die zu maskierende Umgebung geklärt werden:

- (Automatisierte) Ermittlung der anzupassenden Tabellen und Felder.
- Nach der Maskierung müssen die Daten echt aussehen und konsistent sein. (Beispiele: Gültige PLZ, Kreditkartennummer, Kontrollziffer, Namen echt aussehend, richtige Adressen)
- Die Maskierung soll für mehrere Systeme, z.B. in einer Landschaft konsistent sein. (Die meisten Tools betrachten eine Datenbank und genau hier liegt die Schwierigkeit über Datenbanken hinweg die Konsistenz zu gewährleisten)
- Die Daten müssen benutzbar bleiben um sie für gültige Testzyklen einzusetzen. (z.B. statistische Verteilung, Gewichtung, Anzahl der Vorkommnisse, geographische Verteilung)

- Möglichst zeitnahe Anonymisierung.
- Weitestgehende Automatisierung des Ablaufs.
- Das Ergebnis muss die Datenschutzanforderungen erfüllen. (HIPAA, PCI, ITAR…)
- Gesicherter Ablauf der Anonymisierung (Zugriffe sperren, Systeme isolieren, Qualität der Daten)
- Nicht umkehrbar (in 95% der Fälle wird das erwartet)

Szenario
Die Anonymisierung kann für ein System oder für die gesamte Landschaft erfolgen. Auch wenn ein Oft sind komplexe Umgebungen dabei in Betracht zu ziehen wie unterschiedliche Datenbanken oder Applikationen. Es gilt aber weiterhin dass diese Systeme unter sich konsistent sein müssen.

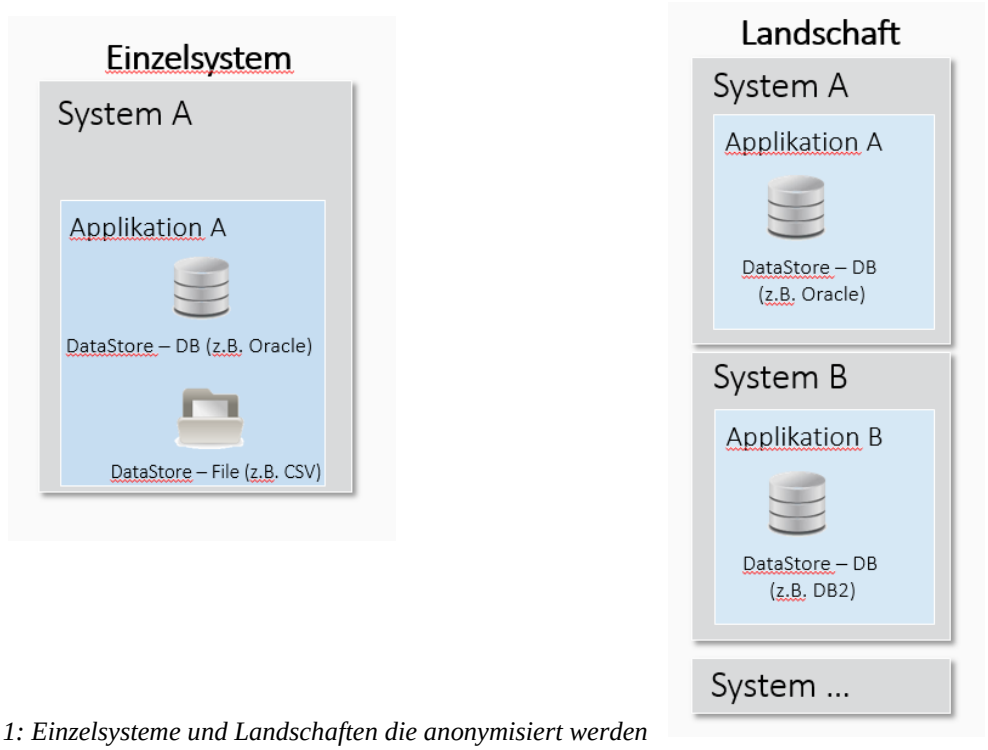
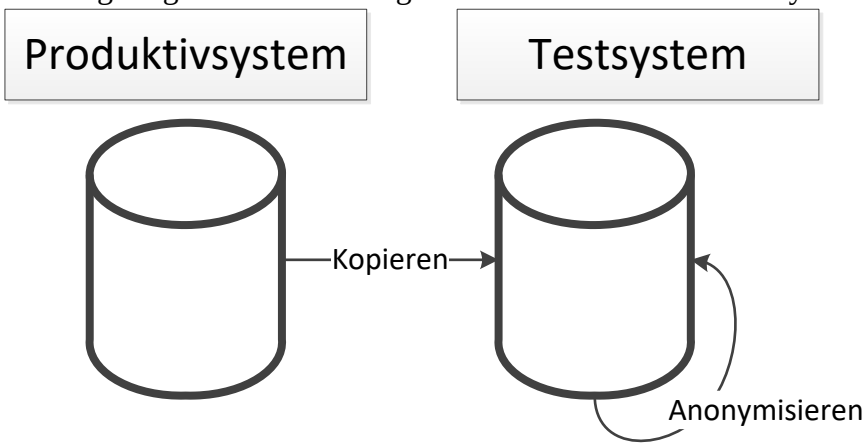


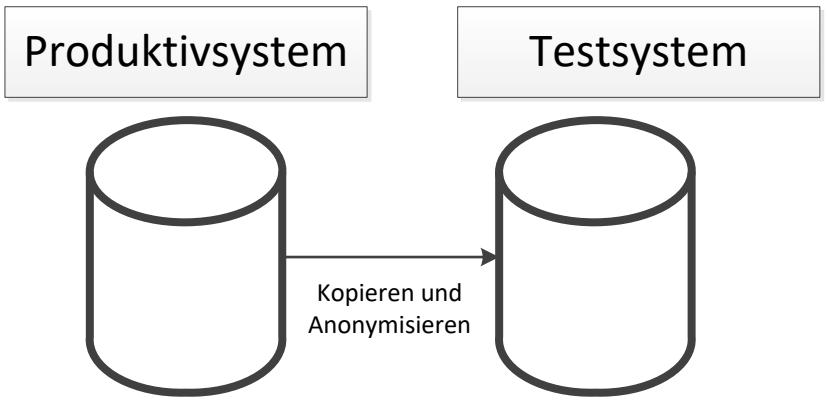
Abb. 1: Einzelsysteme und Landschaften die anonymisiert werden

Weiterhin stellt sich eine weitere Frage. Wo werden Daten anonymisiert?

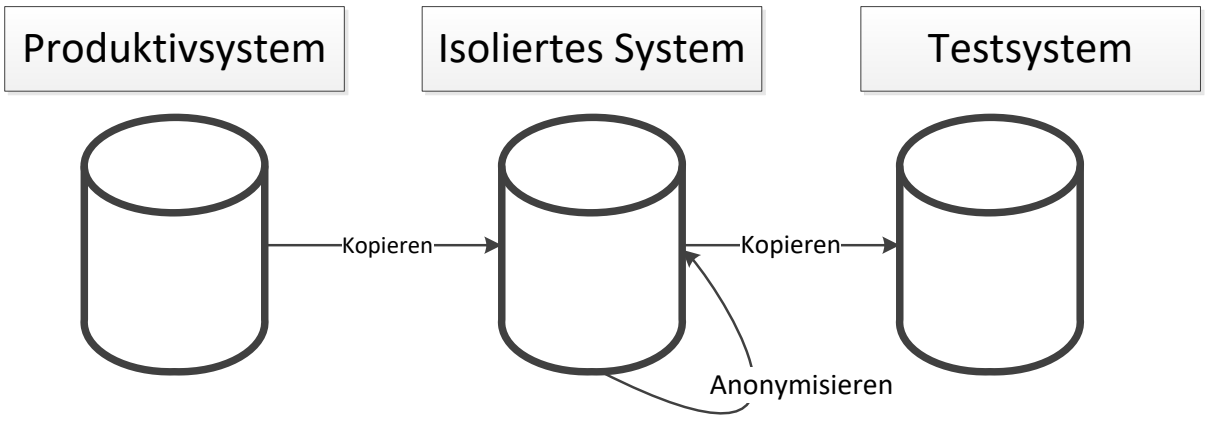
- Wenn produktive Daten bereits im zu anonymisierenden System sind, könnten sie mit geringem Aufwand ausgelesen werden bevor sie anonymisiert werden.



- Eine weitere Möglichkeit ist es die Daten aus dem produktiven System zu lesen, anonymisieren und erst dann in z.B. das Testsystem einzuspielen. Somit wären produktive Daten zu keinem Zeitpunkt im Testsystem. Das Auslesen der Daten aus dem produktiven System kann zu Performanz-Problemen führen, auch die Konsistenz der Daten kann in Frage gestellt werden, da wir davon ausgehen müssen dass sich Daten im Produktiven System ändern.



- Weitere Variante: ein weiteres „isoliertes“ System bereitstellen wo der Backup der produktiven Datenbank eingespielt wird, dieses System ist genau so sicher wie das produktive. Die Daten anschließend in diesem System anonymisieren und erst dann zum Testsystem übertragen z.B. mittels backup/restore der Datenbank. Hier brauchen wir natürlich zusätzliche Ressourcen für so ein isoliertes System.



Ablauf

Die Anonymisierung lässt sich in mehrere Phasen aufteilen.

- Vor der Anonymisierung muss erstmal entschieden werden was anonymisiert werden soll. Z.B. Personaldaten, Adressen usw. Wir können im allgemeinen sagen „die zu anonymisierenden Profile“ bestimmen.

- Damit wir überhaupt anonymisieren können, brauchen wir entsprechende Algorithmen für z.B. Namensanonymisierung, Personalnummer, Adresse usw.

  Es gibt verschiedene Varianten was diese Algorithmen leisten:
  - Inhalt löschen
  - Den Wert auf Konstante setzen
  - Eine Referenztabelle benutzen
  - Shuffle (verwende Daten aus der Datenbank als Referenztabelle)
  - Berechnen (den neuen Wert nach bestimmten Kriterien berechnen z.B. Zufallswerte, Personalnummer neu, Kreditkarte mit Prüfziffer, Sozialversicherungsnummer mit Prüfziffer…)

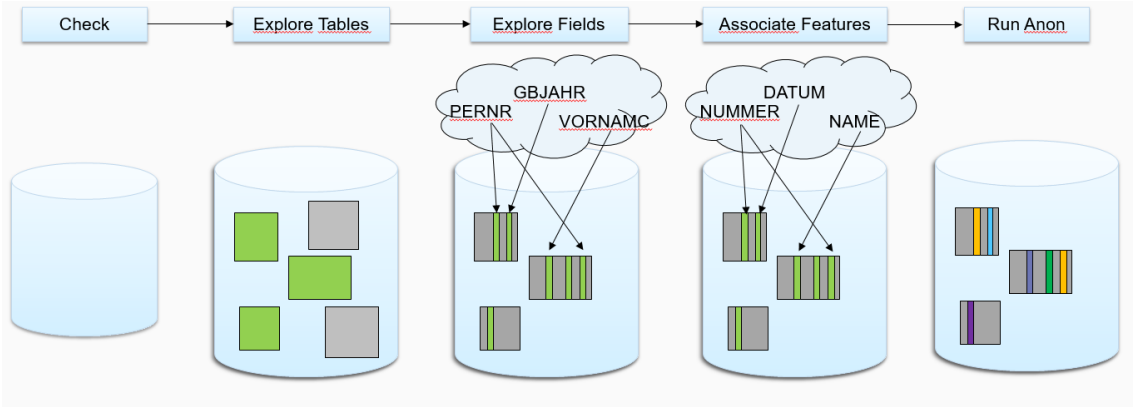
- **WAS** soll anonymisiert werden:
  In der Datenbank müssen Tabellen und die entsprechenden Felder identifiziert bzw. erkannt werden, die zu anonymisieren sind. Dazu kann es bereits eine bekannte Liste von Feldern geben. Eine weitere Variante ist es die Inhalte der Tabellen zu „scannen“ und zu bestimmen welche Felder für die Anonymisierung relevant sind. Ein einfaches Beispiel: ein Feld hat zu 90% als Inhalt Adressen, also dieses Feld ist relevant für die Anonymisierung, weil da die Adressen abgelegt sind. Das Scannen der Daten in einer Datenbank kann zeitintensiv sein.

- **WIE** wird anonymisiert:
  Nun bleiben die zu anonymisierenden Felder mit passenden Algorithmen zu knüpfen.

Somit wissen wir was und wie anonymisiert wird → die Daten können anonymisiert werden.

Der Ablauf würde dann so aussehen:

- CHECK: Prüfe ob alle Voraussetzungen für die Anonymisierung vorhanden sind
- EXPLORE TABLES: Finde Heraus welche Tabellen in Frage kommen.
- EXPLORE FIELDS: Finde heraus welche Felder zu anonymisieren sind (WAS).
- ASSOZIATE FEATURES: Entscheide für Felder mit welchen Algorithmen anonymisiert wird (WIE)
- RUN ANON: Anonymisierung durchführen
- VALIDATE: zum Schluss müssen die Daten validiert werden



Möglichkeiten der Umsetzung

- Eigene Skripte machen (Aufwand aufzusetzen, Überwachung, Wartung)
- Tools die Datenmaskierung durchzuführen
  Optim (IBM)
  Data Masking (Informatica)
  Oracle Data Masking Pack
  Epi-Use
  Libelle Data Masking

Hier gibt es von Libelle ein Framework (**Libelle DataMasking)** welches den Ablauf bereits standardisiert hat und auch während der Durchführung auf saubere Ausführung prüft.

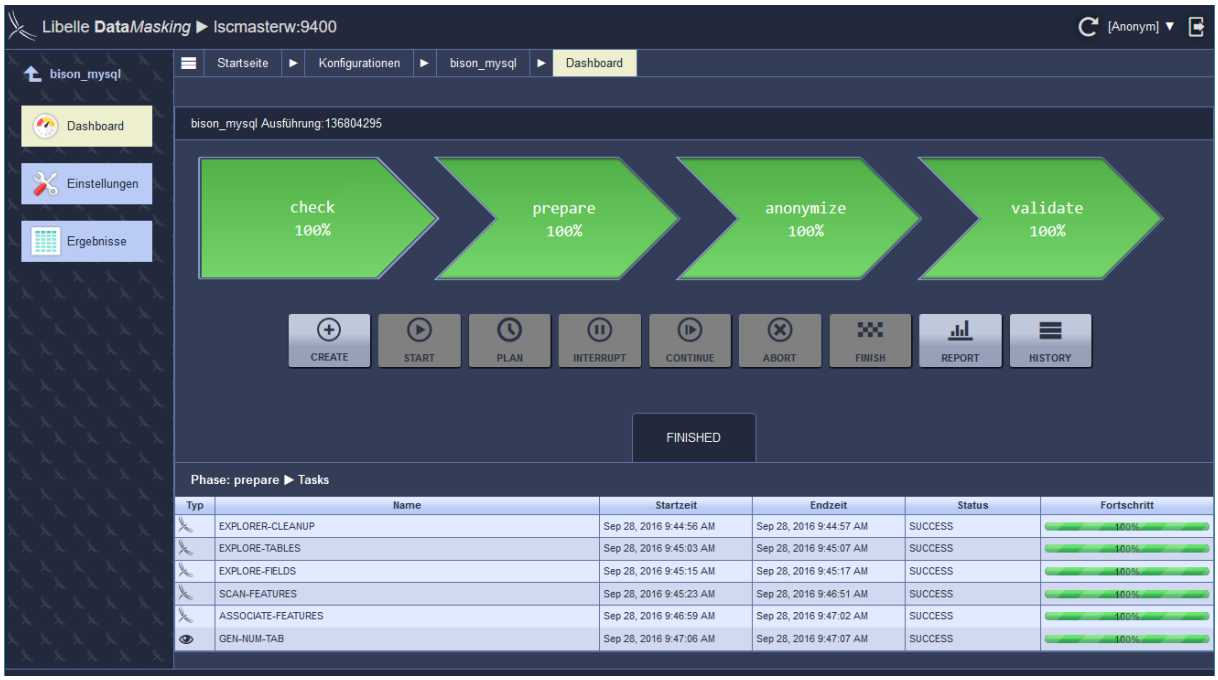
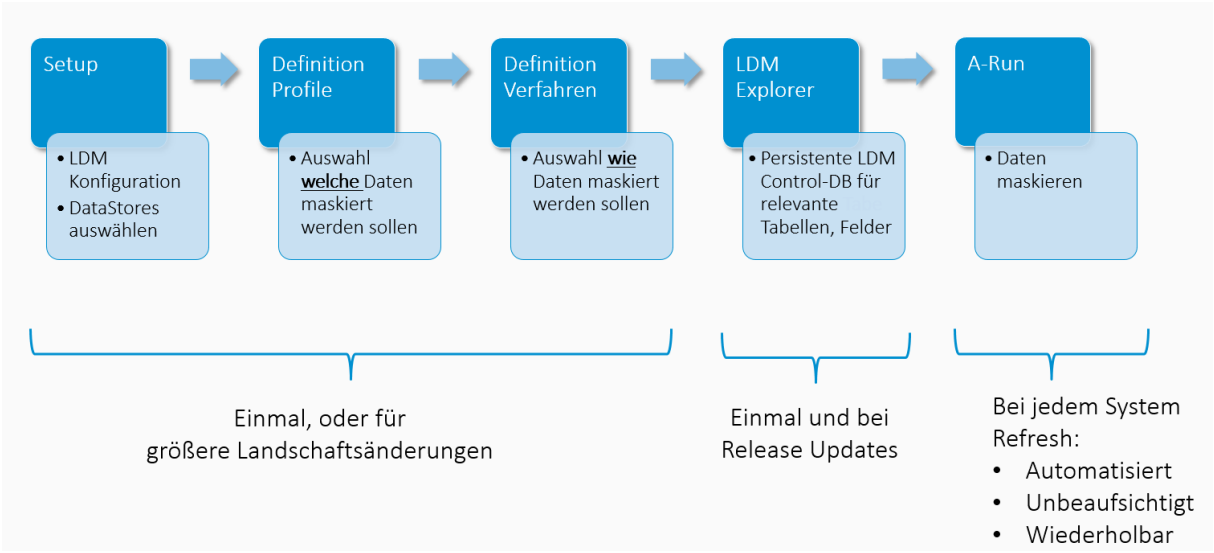


Abb. 1: Libelle Data Masking GUI



Herausforderungen

Organisatorisch: Ansprechpartner der Fachbereiche zusammenbringen und auf gemeinsame Anforderungen einigen, Kompromisse finden.

Vorname, Straße echt, Kreditkartennummern, -> Algorithmen (bild).

Parallelisierung: mehrere Tabellen gleichzeitig (kleine Tabellen)

Partitionierung: eine Tabelle in mehreren Partitionen (große Tabellen)

Abschlußprüfung: Die Schwierigkeit ist es zum einen die Testfälle zu definieren und bei Fehler des Tests zu verifizieren ob es an den Daten oder an dem Testfall liegt.

Datenmaskierung in SAP
Identifizierung der Felder mittels DDICT Tabellen somit ist ein „scannen“ der Datenbank nicht unbedingt notwendig.
Cluster und Pool Tabellen über die SAP Applikation behandeln, was zeitintensiv werden kann.

Zusammenfassung
Jedes Unternehmen hat die Herausforderung der Anonymisierung. Die Lösung soll flexibel sein um künftige Anforderungen abdecken zu können. Es soll auch möglich sein die kundenspezifischen Anforderungen möglichst einfach umzusetzen und zu integrieren.

Kontaktadresse
Miroslav Jakovljevic
Libelle AG
Gewerbestr. 42
D-70565 Stuttgart

Telefon: +49 (0) 711-78335 231
Fax: +49 (0) 711-78335 148
E-Mail mjakovljevic@libelle.com
Internet: www.libelle.de