STABILIT Informatik AG

Synthetische Adressdaten

Agenda

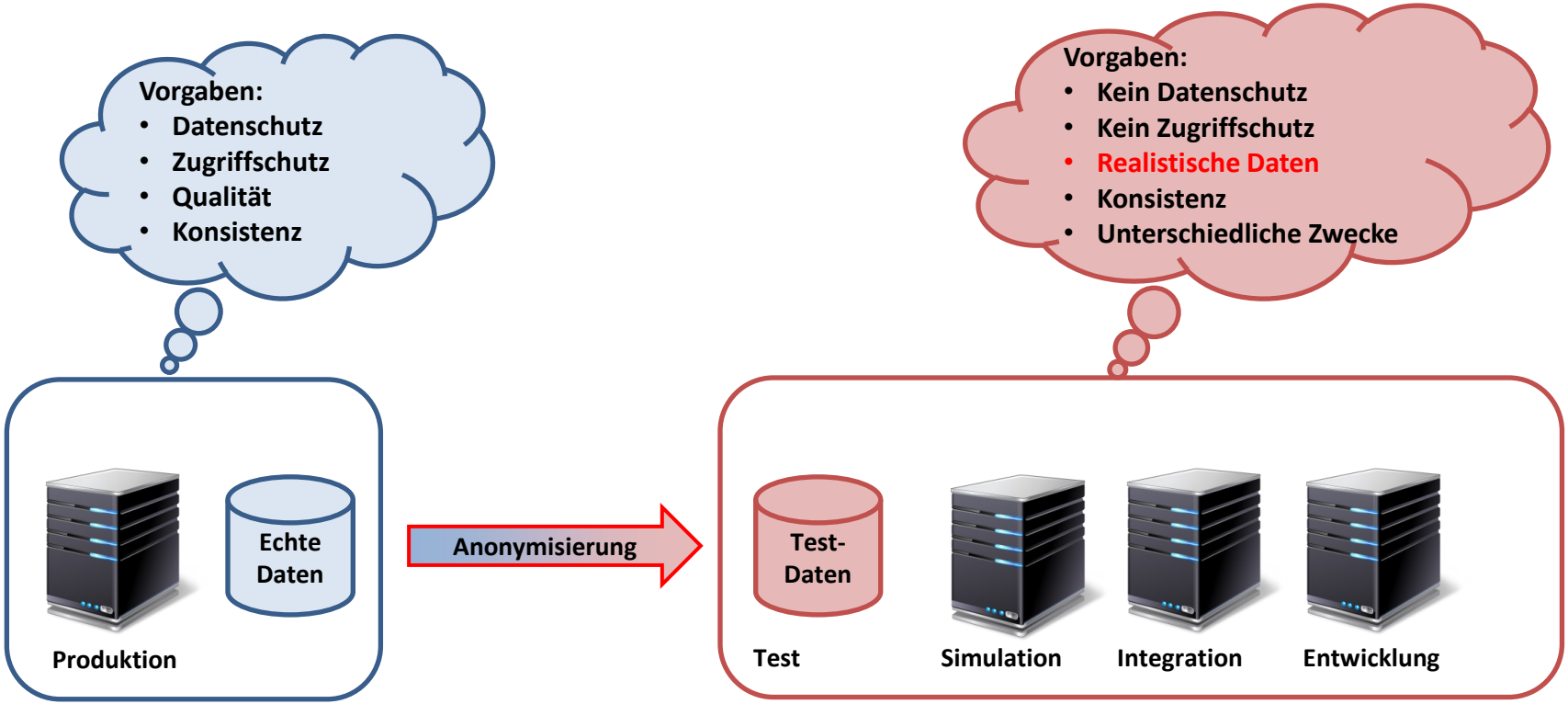
01 Problem

02 Lösung

03 Vorteile

04 Preise

Produktive Daten und Testdaten



Das Problem:

- **Die Anforderungen an die Testdaten sind sehr unterschiedlich**
  - Realistische Daten ⇔ anonyme Daten
  - Unit-Test, Integrations-test, Performance-Test, Regression-Test (Reproduzierbarkeit = Referenzdaten)
  - Kleine Datenmenge ⇔ grosse Datenmenge
- Anonymisierung der echten Daten ist:
  - Kompliziert (viele Abhängigkeiten)
  - Zeitaufwendig
  - hohes Risiko das DSG zu verletzen. (Re-Identifikation)

>>> Steigende Kosten für Bereitstellung der Testdaten

Testdaten müssen realistisch sein

Beispiel Namen:

- F3U2700-CH12 künstlich, viele möglich, kryptisch = unbrauchbar
- Albert Einstein realistisch, brauchbar, leider nicht sehr viele
- Petr Novák realistisch, andere Sprache = unbrauchbar

Was sind realistische Namen?

- Vorname = üblicher Vorname in der Landesregion, Sprachregion
- Nachname = üblicher Name in der Landesregion, Sprachregion

Testdaten müssen realistisch sein

Beispiel Adressen:

• Hagworth 474, 3999 Tupfikon	Künstlich, ungültig =unbrauchbar
• Postweg 474, 5034 Zürich	real, ungültig, bedingt brauchbar
• Postweg 3, 5034 Buchs	real, gültig, brauchbar
• Postweg 10, 5034 Buchs	echte CH Adresse

Was sind realistische Adressen? (Realitätsgrad 1,2,3)

1. Ort und PLZ stimmen, Strasse + Hausnummer beliebig
2. Ort und PLZ und Strasse stimmen, Hausnummer beliebig
3. Ort und PLZ und Strasse und Hausnummer stimmen = echte Adresse

Agenda

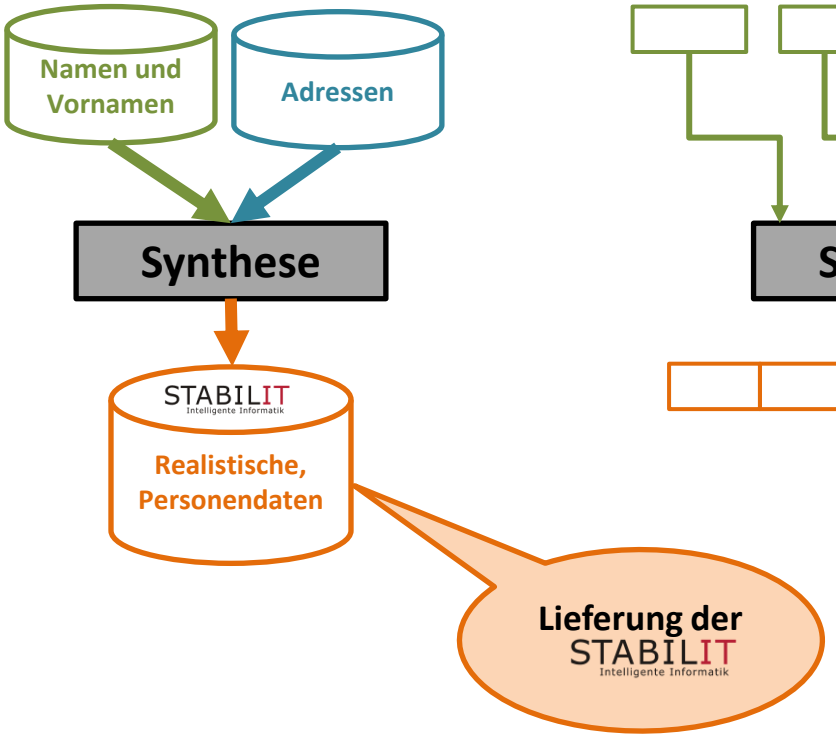
01 Problem

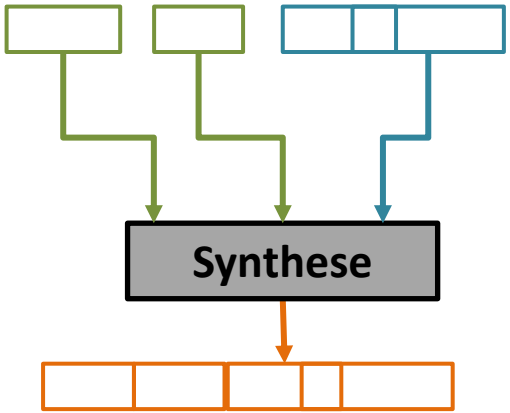
02 Lösung

03 Vorteile

04 Preise

Synthese zu «realistischen» Personendaten





Zufällige Kombination von Name, Vorname und Adresse

Auswahl nach Bedarf:

- Namen nach Sprache (D/I/F)
- Geschlecht (m/w)
- Realitätsgrad der CH-Adresse
- Anzahl Datensätze (10 – 1mio.)

Erstellung der Testdaten mit unseren anonymen Personendaten

Resultat:

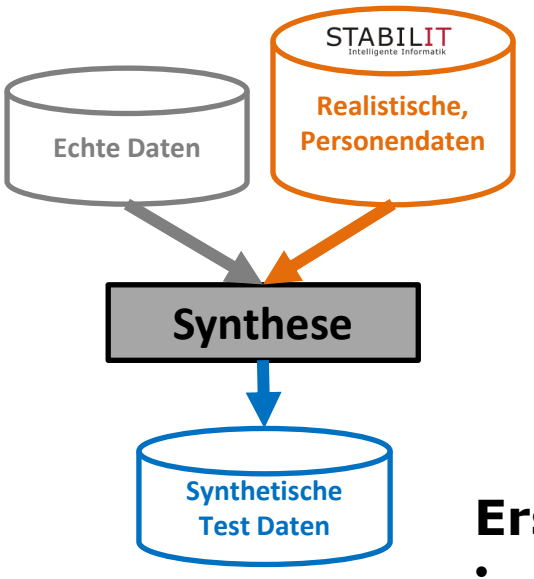
⇒ Anonyme Personendaten
Rückverfolgbarkeit (Re-Identifikation) ist nicht möglich

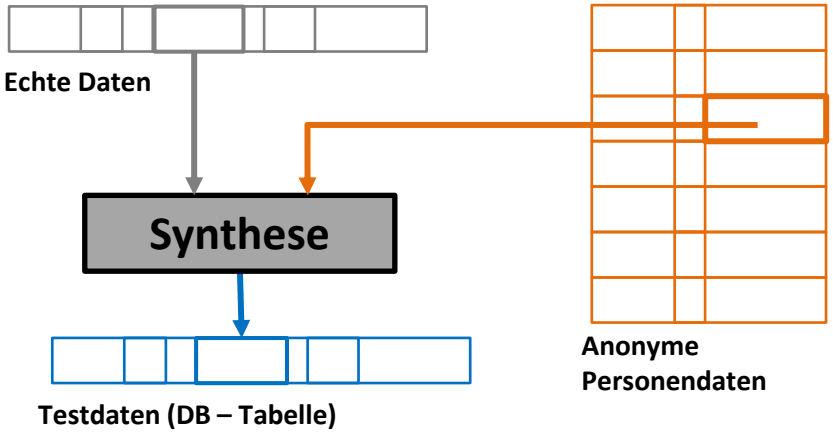
⇒ Realistische Testdaten

⇒ Resultatmenge ist gleich oder kleiner als echte Daten

⇒ Erstellung reproduzierbarer Referenzdaten ist je nach Methode möglich

Einsatz / Verwendung (Variante A)

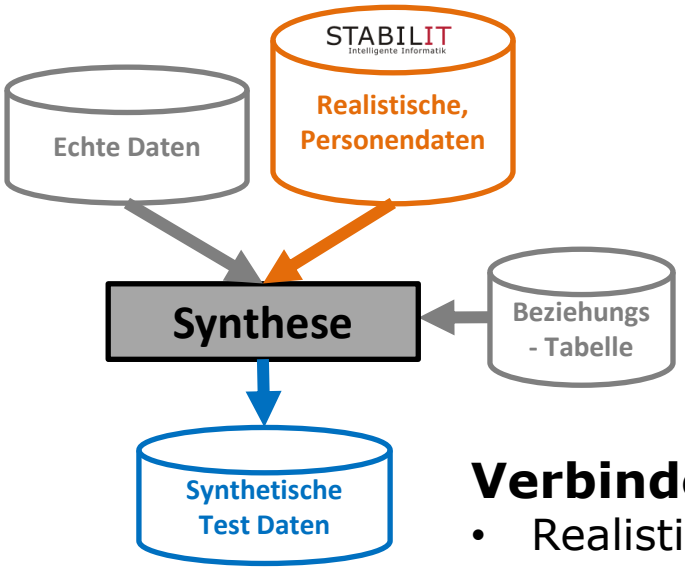


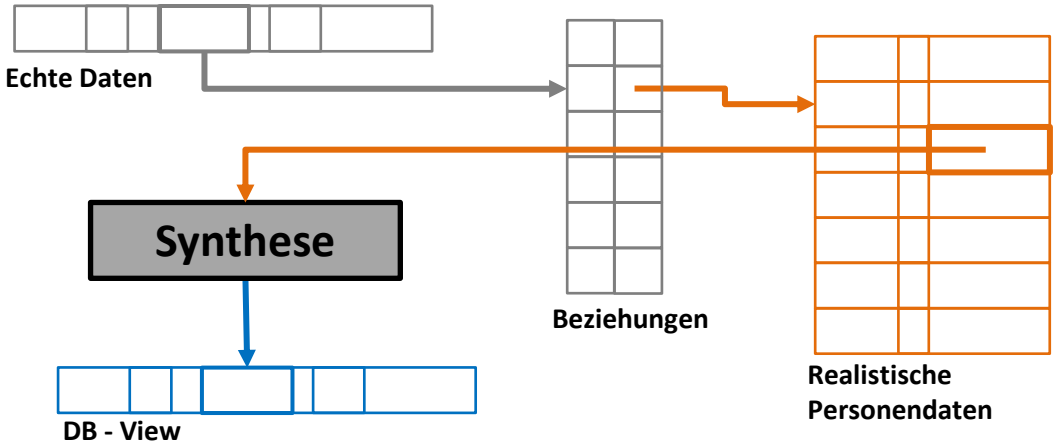


Ersetzen:

- Realistische Vorname + Name + Adresse
- Ersetzen der Attribute in den echten Daten
- Reproduzierbarkeit kann mit Vorsortieren erreicht werden

Einsatz / Verwendung (Variante B)





Verbinden:

- Realistische Vorname + Name + Adresse
- + Beziehungstabelle (Id1, Id2)
- 2 logische Sichten gleichzeitig (echte Daten / Testdaten)
- Reproduzierbarkeit kann mit Vorsortieren erreicht werden

Einsatz / Verwendung

- Weitere Varianten sind möglich
- Synthese ist mit verschiedenen Standard-Tools möglich

- Testdaten für jeden Zweck / Test
- (fast) beliebige Menge an Testdaten
- Erstellung von Referenzdaten ist möglich

⇒ **Einfacher Prozess = niedrige Kosten, schnelle Verarbeitung**

Agenda

01 Problem

02 Lösung

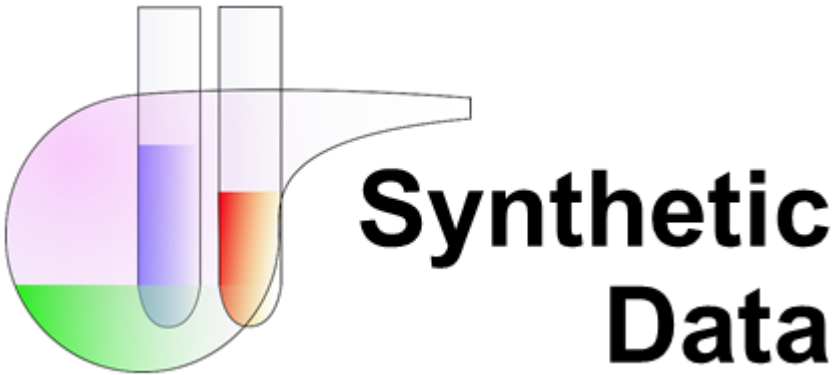
03 Vorteile

04 Preise

Vorteile

- ✓ Anonyme, realistische Testdaten
- ✓ Unterschiedliche Testdaten für unterschiedliche Zwecke
- ✓ Schnelleres, effizientes Verfahren

- ✓ **Niedrige Kosten**



Agenda

01 Problem

02 Lösung

03 Vorteile

04 Preise & technische Spezifikation

Preisliste

- Vorname und Name + Adresse aus der Sprachregion
- Wählbare Parameter:
  - Sprachregion: D,F,I (CZ in Vorbereitung)
  - Geschlecht: m,w
  - Realitätsgrad (1,2,3) der Adresse (Aktualität ist ca. 2009)

Anzahl Datensätze	Preis (CHF) Realitätsgrad 1	Preis (CHF) Realitätsgrad 2	Preis (CHF) Realitätsgrad 3
1’000	1’000	1’500	2’000
10’000	5’000	7’500	10’000
100’000	12’500	18’750	25’000
>100’000	Auf Anfrage	Auf Anfrage	Auf Anfrage

On-line Anfrage: www.stabilit.ch

Technische Spezifikation

- Format = CSV, Encoding = UTF-8
- 1 Headerzeile, Kolonnen sind fest
  - ID (Zahl, 1-n)
  - Vorname (n Zeichen)
  - Name (n Zeichen)
  - Geschlecht (1 Zeichen, m/f)
  - Sprache (code, 5 Zeichen Bsp. «de-ch»)
  - Strasse (n Zeichen)
  - Haus Nr. (Zahl, 1-n)
  - ZIP (Zahl, 4-stellig)
  - Stadt (n Zeichen)
  - Gemeinde (n Zeichen)
  - Kanton / Region (n Zeichen)

Beispiel: Download 10 Datensätze als Gratismuster

ID	Firstname	Lastname	Sex	Language	Street	House-Nr	Zip	Town	County	Canton
1	Maja	Matter	f	de-CH	Obere Allmend	5	4612	Wangen b. Olten	Wangen bei Olten	Solothurn
2	Albano	Ferrari	m	it-CH	via alle Fontane	2	6925	Gentilino	Collina d'Oro	Ticino

Vielen Dank für Ihre Aufmerksamkeit

Motto:

Es kommt nicht darauf an, wie viele Ideen man hat,
sondern darauf, wie viele man umsetzt.