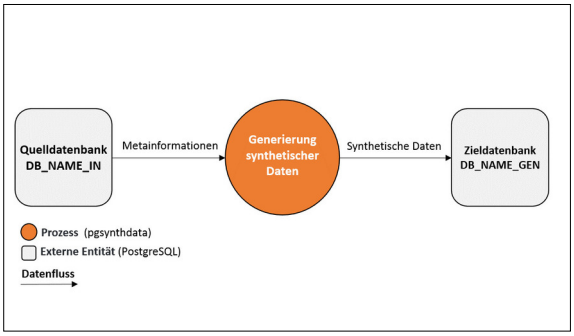
Kevin
Ammann

Student	Kevin Ammann
Examinator	Prof. Stefan F. Keller
Themengebiet	Data Science

Erweiterung eines Generators für rein synthetische Daten

Ein Python-Werkzeug zur Erzeugung von synthetischen Daten für PostgreSQL-Datenbanken

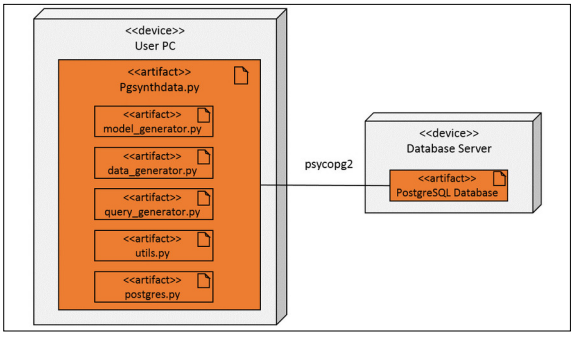


Datenflussdiagramm mit Input und Output des Prozesses von pgsynthdata.
Eigene Darstellung

Data Output | Messages

	tablename name	attname name	most_common_vals anyarray	histogram_bounds anyarray
1	atp_players	country_code	{USA,ESP,AUS,GER,GBR,IT...	{AFG,AHO,AND,AND,ANZ...
2	atp_players	player_id	[null]	{100003,100550,101101,...
3	atp_players	last_name	{Lee,Smith,Kim,Garcia,Rod...	{"A Cantacuzene",Aguirre...
4	atp_players	hand	{U,R,L}	[null]
5	atp_players	first_name	{David,Daniel,Michael,Joh...	{??,Adrien,Alan,Alexandr...
6	atp_players	birth_date	{19860225,19820306,198...	{185903,18920906,1921...

Ausschnitt der statistischen Attributwerte der Tabelle atp_players mit der PostgreSQL-Sicht pg_stats erstellt.
Eigene Darstellung



Verteilungsdiagramm in UML zur Darstellung der Software-Komponenten und -Struktur von pgsynthdata.
Eigene Darstellung

Einleitung: Das Interesse an synthetischen Daten ist in letzter Zeit stark gestiegen, vor allem auch aus den Bereichen maschinelles Lernen sowie Testing, Benchmarking und Tuning von Software und Datenbanksystemen. Um diese Anwendungen zu unterstützen, ist in der vorliegenden Projektarbeit der Datengenerator pgsynthdata - ein Werkzeug zur Erzeugung rein synthetischer Daten - entwickelt worden. Ein erster Prototyp dieses Tools ist bereits unter der Leitung von Professor Stefan Keller am OST Campus Rapperswil entwickelt worden.

In Abbildung 1 ist der Datenfluss von pgsynthdata ersichtlich: Ausgehend von einer originalen PostgreSQL-Quelldatenbank entnimmt der Datengenerator zuerst aus deren Katalog die nötigen Metainformationen. Daraus erzeugt pgsynthdata strukturell und statistisch ähnliche Datensätze und speichert diese in einer neu erzeugten PostgreSQL-Zieldatenbank. Die Synthetisierung übernehmen automatisch konfigurierte Zufallsgeneratoren und Iteratoren. Die Metainformationen ersetzen dabei die ansonsten langwierige manuelle Konfiguration der Zieldatenbank-Struktur und der gewünschten möglichst identischen Datenverteilung.

Ziel der Arbeit: Anregungen von Interessenten und potenziellen Nutzern haben dazu geführt, dass der bestehende Prototyp im Rahmen dieser Arbeit um weitere Features erweitert wurde. Folgende Ziele wurden dabei verfolgt:

- Die Nutzung der statistischen Attributwerte (Most Common Values oder Histogramm) zur Eruierung der Min-/Max-Werte, die Nutzung der erweiterten Statistik-Sicht (pg_stats_ext), sowie die Berücksichtigung von Fremdschlüsselabhängigkeiten sind implementiert.
- Die wesentlichen theoretischen Grundlagen synthetischer Daten sind in einem wissenschaftlichen Kontext wiedergegeben.
- Das Know-how in den Bereichen Datenbanksysteme, PostgreSQL, Python und Git-Repositories ist durch die Auseinandersetzung mit dem Prototyp erweitert.

Ergebnis: Die festgelegten Ziele wurden grösstenteils erreicht und alle Features konnten bis auf eines - die Nutzung der erweiterten Statistik-Sicht - erfolgreich umgesetzt werden. Um die Unter- sowie Obergrenzen der synthetischen Daten festzulegen, werden nun die Min- und Max-Werte der jeweiligen Attribute von den Most Common Values oder von den Histogramm-Grenzen ausgelesen. Die zweite Abbildung veranschaulicht diese beiden statistischen Werte am Beispiel der Tabelle atp_players. Mithilfe einer vorgängigen topologischen Sortierung der Tabellen können neu auch Datenbanken mit Fremdschlüsselabhängigkeiten synthetisiert werden, sofern sie in der ersten Normalform vorliegen. Damit sich der Prototyp auch künftig besser erweitern lässt, wurde nebst der Implementation automatisierter Tests zusätzlich ein Refactoring durchgeführt. Abbildung 3 veranschaulicht den Aufbau der Software pgsythdata als Resultat des Refactorings. Schliesslich wurde noch das Ergebnis analysiert, indem die Ausführungspläne ausgewählter SQL-Abfragen auf die generierte Zieldatenbank mit denjenigen der originalen Quelldatenbank verglichen wurden.