Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

Анонимизација на оперативна база на податоци

Ментор: Ефтим Здравевски

Климентина Џепароска 161525

Грација Николовска 161520

Милена Ѓорѓиевска Перушеска 161536

За произволна база, во овој проект AdventureWorks2012, се генерираат SELECT наредби кои ги маскираат колоните кои се означени како Лични или Заштитени (пр. матичен број, име и презиме, состојба на трансакциска сметка, итн.). Некои од нив се маскираат со други вредности, некои се изоставуваат, итн.. Се посветува и внимание на безбедносни мерки како и протоколи за заштита на база на податоци. (стандарди, мерки, практики). Визуелниот приказ на базата се наоѓа на линкот.

=> Анонимизација. Типови на анонимизација:

- **None** - останува оригиналната вредност
- **Omit** - се отстранува колоната од SELECT наредбата (ќе има помалку колони во излезниот поглед кој одговара на табелата)
- **Default value** - вредноста што била доделена во default constraint или проследена во конфигурацијата на анонимизацијата
- **Set Null** - поставува null вредност во таа колона. Ако колоната не е nullable, уште на почеток треба да се врати соодветна грешка
- **Random** - случајна вредност, со тоа што два пати ако се повика за иста вредност се добива различна анонимизирана вредност(пр. 1 -> 123, вториот пат 1 -> 234)
- **RandomPseudonym** - случајна вредност, со тоа што два пати ако се повика за иста вредност (пр. 1 -> 123, вториот пат исто така 1 -> 123). За ова треба да се чува дополнителна табела со мапирањата
- **RandomFromSet** - да се одбере случајна вредност од множество на вредности. На пример, нека колоната TransactionTypeId nvarchar(20) е надворешен клуч кон TransactionType(Id) со можни вредности: pending, processing, cancelled, finished, error. Тогаш со овој тип на анонимизација на TransactionTypeId треба да се додели една од дозволените вредности дефинирани од надоврешниот клуч. Значи овие вредности не треба да се проследуваат при конфигурирањето, туку да се земаат динамички од референцираната табела. При 2 маприања на ист статус, може да се добија различни мапирани вредности (пр. cancelled еднаш го мапира во finished, втор пат во pending, итн.)
- **RandomPseudonymFromSet** - исто како претходното, со тоа што треба секогаш истата вредност да се мапира во иста вредност (пр. cancelled секогаш го мапира во finished)

Проектот се содржи од два дела: Python & MSSQL. Кодот е достапен на линкот. Во двата дела како влез се користат табелите кои што ќе се анонимизираат, како и врските меѓу нив, проследени во продолжение:

Table_Name	Column_Name	pe {nullable,	Anonymization_Type {None, Omit, Randomize, Default value, Encode}
HR.Employee	NationalIDNumber	unique	RandomPseudonym
HR.EmployeePayHistory	Rate	not null	Omit
Sales.SalesOrderHeader	CreditCardApprovalCode	nullable	Null
Sales.CreditCard	CardNumber	unique	Random
Person.EmailAddress	EmailAddress	nullable	None
Person.Password	PasswordHash	not null	RandomPseudonym
Purchasing.Vendor	ModifiedDate	not null	Default
Purchasing.PurchaseOrderHeader	EmployeeID	not null	RandomFromSet
Production.Product	ProductId	unique	RandomPseudonymFromSet
Production.TransactionHistory	TransactionID	not null	Omit

Tables.csv - CSV документ со име на табела, име на колона, тип на податоци, nullable, unique, Anonymization Type (None, Omit, Randomize, Default value, Encode).

Ref1_Table_Name	Ref1_Colum_Name	Ref2_Table_Name	Ref2_Colum_Name
Purchasing.PurchaseOrderHeader	EmployeeID	HumanResources.Employee	BusinessEntityID
Production.TransactionHistory	ProductId	Production.Product	ProductId

Keys.csv - CSV документ со информации за надворешните клучеви со вкупно 4 колони: име на референцирачка табела, име на референцирачка колона, име на референцирана табела, име на референцирана колона. Првите две колони како пар дефинираат од каде е надворешниот клуч, вторите две дефинираат до што покажува надворешниот клуч. Ќе се ограничиме на прости надворешни клучеви составени само од една колона.

- **Python**

Влез: Вчитување на информации за базата на податоци во вид на две датотеки

=> Input(tables.csv, keys.csv)

=> Во зависност од типот на анонимизација, се повикуваат соодветните функции.

Излез:

=> Се генерираат .CSV датотеки кои што поминале низ процес на анонимизација, во кои што се оддржува конзистентност во мапирањето кај типовите на анонимизација.

=> Преку .CVS се генерираат погледите за анонимизација. За секоја табела има соодветен поглед.

- **MSSQL**

Влез: Вчитување на базата на податоци

=> exec Data_Anonymization @table='table_name', @column='column_name'',@type='type_name'

=> Во зависност од типот на анонимизација, се повикуваат соодветните процедури од главната процедура.

Излез:

=> Се генерираат табели кои што поминале низ процес на анонимизација во кои што се оддржува конзистентност во мапирањето кај типовите на анонимизација.

=> Преку SELECT се генерираат погледите за анонимизација. За секоја табела има соодветен поглед.

=> Се верификува дека нема синтаксички и логички грешки.

Маскирањето, односно анонимизацијата на податоците е општ метод за давање на автентични делови или на сите автентични парчиња на податоци на начин што ги заштитува вистинските податоци да не бидат целосно прикажани, а може да се користат разни техники за да се воспостави маска за податоци. Маските со податоци може да бидат полни (сокривање на сите оригинални знаци на податоци) или делумни (замаглување само некои од знаците на податоците). Маскирањето на податоците е исто така клучен услов во согласност со стандардите во индустријатаи регулативите за заштита на личните податоци и приватноста.