УДК 004.658.2

ОБЕСПЕЧЕНИЕ ЦЕЛОСТНОСТИ ВНЕШНИХ КЛЮЧЕЙ МАСКИРОВАННОЙ БАЗЫ ДАННЫХ

Михаил Коломыцев, Светлана Носок, Анастасия Мазуренко

Статья посвящена актуальной проблеме защите информации в базах данных. Авторы рассматривают метод защиты данных путем маскирования. Суть маскирования данных состоит в необратимой замене в базе данных конфиденциальной информации (например, данных, идентифицирующих конкретных людей) несекретными данными для предотвращения доступа к ней неавторизованными пользователями. Как правило, конфиденциальные данные заменяются похожими на реальные значения, чтобы их можно было использовать в тестовых системах с гарантией, что первоначальные данные не могут быть получены, извлечены или восстановлены. Маскирование позволяет владельцам базы данных самим определять, сколько конфиденциальных данных следует отображать, при минимальном влиянии на работу приложений – данные должны оставаться функционально пригодными для прикладной обработки (в основном, задач тестирования, обучения и т.п.). В данной статье авторы предлагают методику маскирования первичного ключа базы данных (БД). Данная методика реализована в виде хранимой процедуры для системы управления базами данных (СУБД) MS SQL Server.

Ключевые слова: *база данных, защита персональных данных, маскирование данных, конфиденциальные данные, информационная система.*

Актуальность и постановка задачи маскирования данных.

Практика разработки информационных систем показывает, что кроме производственной (основной) базы данных возникает задача создания ее копий, непроизводственных (тестовых) баз данных. Связано это с рядом причин:

– необходимостью проведения работ по развитию информационной системы, ее тестированию, обучению пользователей;

– копии БД могут использоваться для аналитической обработки (data mining);

– предоставление данных партнерам в рамках производственной или торговой кооперации и других целей.

По некоторым оценкам, число непроизводственных баз данных может достигать 6-8 экземпляров. И, если в производственной БД поддерживается адекватная политика безопасности, то в тестовых БД вопросам безопасности внимания уделяется меньше, что повышает риски утечки данных. Типичными случаями является бесконтрольное создание новых учетных записей в тестовой БД и извлечение из нее данных для других приложений.

Согласно данных Oracle [1], почти половина компаний, принимавших участие в проведенном обследованиях, заявляли, что не контролируют возможность потери данных из тестовых БД.

Особую актуальность в Украине приобрел вопрос защиты персональных данных, после принятия соответствующего закона [2]. Согласно закону, владельцы баз персональных данных обязаны обеспечить их защиту.

Для защиты данных в такой ситуации можно использовать подход, который называется маскирование данных [3]. Маскирование данных представляет собой процесс обезличивания или сокрытия определенных данных в таблицах базы. Маскирование данных по своей сути защищает конфиденциальные данные от неавторизованного доступа, изменяя их значения, но сохраняя при этом первоначальные ограничения на значения данных.

Важной особенностью непроизводственных БД является то, что они должны быть функционально эквивалентными производственной БД. Это означает, что для обработки информации в тестовой БД можно использовать те же приложения, что и для основной БД.

Постановка задачи. Рассмотрим ситуацию, когда существует основная таблица и несколько связанных с ней подчиненных таблиц. Ссылка на первичный ключ основной таблицы является внешним ключом в подчиненной. Необходимо маскировать значения первичного ключа, сохранив связи с подчиненными таблицами.

В связи с этим целью данной работы является разработка методики статического маскирования данных методом подстановки, обеспечивающей маскирование первичного ключа и целостность внешних ключей БД. Данная методика реализована в виде хранимой процедуры для СУБД MS SQL Server.

Требования к методам маскирования.

К типичным ограничениям процесса маскирования данных можно отнести [3,5]:

Невозможность восстановления исходных данных. Это очевидное условие, без которого маскирование теряет смысл. При реализации процесса маскирования необходимо учитывать, что результат преобразования не должен быть обратимым в том смысле, что у лиц, не имеющих доступа к ключевой информации, не должно быть возможности восстановить оригинальный текст, используя маскированный.

Сохранение принадлежности данных к определенному домену. Данное ограничение можно представить в виде набора частных требований, например:

Сохранение типа данных. Результаты маскирования должны относиться к тем же базовым типам данных, что и исходные данные.

Сохранение формата. Маскированные данные должны иметь такую же структуру, что и исходные данные. Это означает, что, если исходные данные имеют, например, размер от 2 до 30 символов, маскированные данные должны также отвечать этому условию. Типичным примером является маскирование даты, которое должно происходить в правильных диапазонах для дня, месяца и года. Это означает, что алгоритм маскирования должен определить «смысл» исходных данных, таких как «31.03.2013», «31 марта 2013», и «03.31.2013», а также генерировать подходящую дату в том же формате.

Уникальность. Каждому уникальному исходному значению должно соответствовать уникальное маскированное значение. Это особенно важно для поддержания целостности первичных и внешних ключей (обеспечение целостности сущностей и ссылочной целостности).

Ссылочная целостность. Маскированные данные не должны нарушать ссылочную целостность таблиц БД. Например, если ИНН является первичным ключом, то все маскированные экземпляры одного и того же значения ИНН в связанных таблицах должны быть одинаковыми.

Семантическая целостность. Исходя из семантики предметной области, на значения отдельных полей таблицы с помощью условия *check* накладываются ограничения. Для поддержания бизнес-процессов в информационной системе, маскированные значения должны отвечать таким же ограничениям. Например, необходимость сохранения гендерных признаков (при замене имен мужские имена должны заменяться на мужские, а женские – на женские).

Применимость алгоритма маскирования ко всему множеству данных домена. Если, например, маскируются данные технологического процесса, то алгоритм маскирования должен адекватно обрабатывать весь диапазон исходных данных.

Маскирование должно быть **автоматизированным, легко повторяемым процессом**. В случае, если данные в производственной БД меняются часто, маскирование должно быть простой и эффективной частью процесса создания непроизводственных БД.

Кроме основных требований, могут возникать дополнительные, обусловленные особенностями предметной области, например:

Сохранение обобщенных значений. Суммарные и средние значения по маскированной колонке таблицы должны совпадать с оригинальными (в точности, либо с определенным допуском).

Статистическое распределение значений. В некоторых случаях важно сохранить информацию о таких статистических характеристиках, как характер распределения. Например, если в БД содержится информация о географическом распределении онкобольных по почтовым индексам, то произвольная замена почтовых индексов может исказить результаты анализа.

Методы маскирования.

Используются следующие методы маскирования [3,4]:

Подстановка. В этом случае используется замена одного значения другим. Например, фамилия субъекта заменяется случайно выбранной фамилией из таблицы подстановок, созданной на основе телефонного справочника.

Сложность возникает, если необходимо сгенерировать очень большую таблицу подстановок. Например, если необходимо сгенерировать несколько миллионов реалистичных почтовых адресов клиентов.

Замена на символ-константу. Данный метод является частным случаем метода подстановки, когда все маскируемые символы заменяются одним и тем же символом, например, «X». В этом случае маскированный телефонный номер будет иметь вид «(XXX) XXX-XXXX». Это самый простой и быстрый метод маскирования, однако его ценность невелика.

Перестановка (shuffling). Перестановка является методом рандомизации существующих значений вертикально в наборе данных, т.е. перестановка в столбце таблицы. Однако, если применяется только перестановка, то маскирование является ненадежным. Лицо, обладающее какими–

либо знаниями о реальных значениях, может последовательно восстановить исходные данные.

Метод перестановок эффективен при большом количестве исходных данных. Поскольку отпадает необходимость генерации новых значений, метод является простым и достаточно быстрым. При его реализации необходимо особое внимание уделить рандомизации процесса перестановки.

Размывание (bluring). Оригинальное значение заменяется случайным, но близким (в рамках определенного диапазона) значением. Например, замена данных реальных продаж на произвольное значение, которое отличается в диапазоне 5% от оригинального. Метод может быть полезен, если необходимо скрыть корреляцию между исходными числовыми значениями.

Усреднение. В этом методе оригинальные числа заменяются случайными таким образом, что среднее значение по всему набору маскированных значений остается таким же, как и в оригинальном наборе.

Де-идентификация. Общее название для методов, позволяющих таким образом преобразовать исходную информацию, идентифицирующую личность, чтобы исчезла связь с данной личностью. Де-идентификация используется для маскирования сложных наборов данных, охватывающих несколько столбцов таблицы БД [7].

Разбиения на лексемы (токены). В этом методе элементы данных заменяются случайными заполнителями (токенами). Представление данных в виде токенов является необратимым, так как токен логически не связан с первоначальным значением.

Шифрование с сохранением формата. В этом методе маскирования данные преобразовываются в зашифрованную форму таким образом, что общий вид оригинального значения сохраняется.

По способу организации процесса маскирования различают статическое и динамическое маскирование.

При статическом маскировании создается копия производственной БД с заменой защищаемых данных на маскированные значения. Статическое маскирование позволяет создать реалистичные тестовые БД и снизить риск раскрытия информации в непроизводственной среде, поскольку тестовая БД не содержит немаскированных данных.

Динамическое маскирование выполняется над производственной БД. Процесс преобразования осуществляется в промежуточном программном слое, между производственной БД и приложением. Динамическое маскирование срабатывает в момент обращения к БД и модифицирует ее ответы таким образом, что выдаются обезличенные данные. Решения по динамическому маскированию предназначены для защиты данных, хранящихся в производственных БД. Такой подход позволяет снизить риск утечки данных от действий инсайдеров.

В некоторых СУБД (например, Oracle [1], MS SQL Azure [6], IBM [8]) средства маскирования данных уже реализованы. Поскольку коммерческие решения не всегда доступны, разрабатываются различные частные методики маскирования.

Обеспечение целостности внешних ключей маскированной БД в MS SQL Server.

Методика маскирования состоит из следующих этапов:

- В основную таблицу добавляется колонка с признаком *identity*. Данная колонка содержит маскированные значения первичного ключа. Свойство *identity* означает, что исходные значения первичного ключа будут заменены возрастающей последовательностью целых чисел.
- Разрывается связь между основной таблицей и подчиненными. В подчиненных таблицах в колонку внешних ключей заносятся значения из колонки с признаком *identity* основной таблицы.
- В основной таблице значения первичного ключа заменяются значениями из колонки с признаком *identity*.
- Восстанавливаются внешние ключи в связанных таблицах.
- Колонка с признаком *identity* удаляется.

Ниже приводится пример хранимой процедуры, реализующей данную методику. В процедуре используются обращения к системным таблицам и функциям:

– ***sys.sysforeignkeys***. Системная таблица с данными о внешних ключах. Содержит колонки:

– *fkeyid* (идентификатор таблицы с ограничением FOREIGN KEY),

– *rkeyid* (идентификатор таблицы, на которую ссылается FOREIGN KEY),

– *fkey* (идентификатор колонки в которой содержится ссылка),

– *rkey* (идентификатор колонки на которую указывает ссылка).

– ***syscolumns***. Системная таблица. Каждая строка таблицы содержит описание каждой колонки всех таблиц БД. Содержит колонки:

– *id* (идентификатор таблицы),

– *colid* (идентификатор колонки),

– *name* (название колонки).

– ***sys.objects***. Системная таблица с описанием всех объектов БД. Каждая строка таблицы содержит описание отдельного объекта БД. Содержит колонки:

– *object_id* (идентификатор объекта),

– *schema_id* (идентификатор схемы).

– ***object_name***. Встроенная функция. Возвращает имя объекта схемы по заданному идентификатору объекта.

Текст процедуры:

```
CREATE procedure SP_REFERENCES_MASK(@TBL_NAME VARCHAR(256),@PK_COL_NAME VARCHAR(256),@IDENTITY_COL VARCHAR(256)='MASK_ID',@MASK_WORD VARCHAR(256)='MASKED')
AS
SET NOCOUNT ON
declare @Sqlstring VARCHAR(MAX);
declare @t table(SCH_ID int,Cname sysname, Bname sysname, fkeyid int, constid int, rkeyid int)
/* 1. Создать временную таблицу с информацией о внешних ключах */
insert into @t
SELECT SOF.schema_id AS SCH_ID, C.Name as CName, B.Name as Bname, A.fkeyid, A.constid, A.rkeyid
FROM sys.sysforeignkeys AS A INNER JOIN
 sys.syscolumns AS b ON A.fkeyid = b.id AND A.fkey = b.colid INNER JOIN
 sys.syscolumns AS c ON A.rkeyid = c.id AND A.rkey = c.colid INNER JOIN
 sys.objects AS SOF ON SOF.object_id = A.fkeyid
where SCHEMA_NAME(SOF.schema_id) +'.'+ object_name(A.rkeyid) = SCHEMA_NAME(SOF.schema_id) +'.'+ @TBL_NAME and C.name = @PK_COL_NAME
/* 2. Добавить колонку с признаком Identity */
Set @Sqlstring = N'ALTER TABLE '+ @TBL_NAME + ' add ' + @IDENTITY_COL +' INT IDENTITY(1,1)'
EXEC (@Sqlstring)
/* 3. удалить все ссылки на первичный ключ в связанных таблицах */
select distinct
 @Sqlstring = N'ALTER TABLE ' + SCHEMA_NAME(SCH_ID) +'.'+ object_name(fkeyid) +' DROP CONSTRAINT ' + object_name(constid)
from @t
EXEC (@Sqlstring)
/* 4. Установить в колонке внешнего ключа значения из колонки identity основной таблицы */
select
@Sqlstring = N'UPDATE ' + SCHEMA_NAME(SCH_ID) +'.'+ object_name(fkeyid) +' SET ' + Bname + ' = CAST('+object_name(rkeyid)
+ '.'+ @IDENTITY_COL + ' AS VARCHAR(9))' +' + '''+ @MASK_WORD +'''' +' FROM ' + SCHEMA_NAME(SCH_ID) +'.'+ object_name(fkeyid)+' '
+ 'JOIN ' + SCHEMA_NAME(SCH_ID) +'.'+ object_name(rkeyid) + ' '+
+ 'ON ' + SCHEMA_NAME(SCH_ID) +'.'+ object_name(rkeyid) +'.' +CNAME +' = '
+ SCHEMA_NAME(SCH_ID) +'.'+ object_name(fkeyid) + '.'+ Bname
from @t
EXEC (@Sqlstring)
/* 5. Внести в колонку первичного ключа основной таблицы значения из колонки identity */
SELECT @Sqlstring = N'UPDATE '+ @TBL_NAME + ' SET ' + @PK_COL_NAME + ' = CAST('+ @IDENTITY_COL + ' AS VARCHAR(9))' +'+'''+ @MASK_WORD+''''
EXEC (@Sqlstring)
```

```
/* 6. Восстановить удаленные в п.2 внешние ключи, т.е. восстановить ссылки
 из подчиненных таблиц на первичный ключ основной таблицы */
select
@Sqlstring = N'ALTER TABLE ' + SCHEMA_NAME(SCH_ID) +'.'+ object_name(fkeyid) +' ADD
CONSTRAINT ' + object_name(constid)
+' FOREIGN KEY (' + Bname + ') REFERENCES ' + SCHEMA_NAME(SCH_ID) +'.'+ ob-
ject_name(rkeyid) + '(' + Cname + ')'
from @t
EXEC (@Sqlstring)
/* 7. Удалить колонку с признаком Identity из основной таблицы */
Select @Sqlstring = N'ALTER TABLE '+ @TBL_NAME + ' DROP COLUMN ' + @IDEN-
TITY_COL
EXEC (@Sqlstring)
GO
```

Пример вызова процедуры:

SP_References_Mask Person, INN

Заключение. Предложенная авторами методика использует при маскировании данных информацию о связях между таблицами БД, что позволило обеспечить ссылочную целостность таблиц после преобразования. Необратимая замена первичных ключей в процессе маскирования гарантирует конфиденциальность данных. Преобразованные по изложенной методике данные могут использоваться в непроизводственных БД для таких целей, как тестирования приложений и обучения персонала.

ЛИТЕРАТУРА

[1]. Ahmed W. Data Masking Best Practice [Электронный ресурс] / W. Ahmed, J. Athreya. – 2013. – Режим доступа: http://www.oracle.com/us/products/database/data–masking–best–practices–161213.pdf.

[2]. Закон України «Про захист персональних даних» від 20.12.2012 №2297– VI.

[3]. Коломыцев М.В., Южаков А.М. Защита персональных данных методом маскирования / М. В. Коломыцев, А. М. Южаков // Захист інформації. – 2013. – Т. 15, № 4. – С.382-387. – Режим доступа: http://nbuv.gov.ua/j–pdf/Zi_2013_15_4_17.pdf

[4]. Understanding and Selecting Data Masking Solutions: Creating Secure and Useful Data [Электронный ресурс]. – 2012. – Режим доступа: https://securosis.com/assets/library/reports/UnderstandingMasking_FinalMaster_V3.pdf.

[5]. The Five Laws Of Data Masking [Электронный ресурс]. – 2008. – Режим доступа: https://securosis.com/blog/the–five–laws–of–data–masking.

[6]. Get started with SQL Database Dynamic Data Masking (Azure portal) [Электронный ресурс] – Режим доступа: https://azure.microsoft.com/en–us/documentation/articles/sql–database–dynamic–data–masking–get–started–portal/.

[7]. A Globally Optimal k–Anonymity Method for the De–Identification of Health Data [Электронный ресурс] / [K. El Emam, F. Dankar, R. Issa и др.]. – 2009. – Режим доступа: http://jamia.oxfordjournals.org/content/16/5/670.full.

[8]. Haldeman J. Compare IBM data masking solutions: InfoSphere Optim and DataStage [Электронный ресурс] / John Haldeman. – 2012. – Режим доступа: http://www.ibm.com/developerworks/data/library/techarticle/dm–1211maskingsolution/.

REFERENCES

[1]. Ahmed W. Data Masking Best Practice [Electronic resource] / W. Ahmed, J. Athreya. – 2013. – Access mode: http://www.oracle.com/us/products/database/data–masking–best–practices–161213.pdf.

[2]. Personal Data Protection Act of Ukraine No. 2297–VI, 20.12.2012.

[3]. Kolomytsev M.V., Yuzhakov A.M. Protection of personal data by masking. / M.V. Kolomytsev, A.M. Yuzhakov // Ukrainian Information Security Research Journal. – 2013. – Vol. 15, № 4, pp. 382–387. – Mode access: http://nbuv.gov.ua/j–pdf/Zi_2013_15_4_17.pdf

[4]. Understanding and Selecting Data Masking Solutions: Creating Secure and Useful Data [Electronic resource]. – 2012. – Mode access: https://securosis.com/assets/library/reports/UnderstandingMasking_FinalMaster_V3.pdf.

[5]. The Five Laws Of Data Masking [Electronic resource]. – 2008. – Mode access: https://securosis.com/blog/the–five–laws–of–data–masking.

[6]. Get started with SQL Database Dynamic Data Masking (Azure portal) [Electronic resource] – Mode access: https://azure.microsoft.com/en–us/documentation/articles/sql–database–dynamic–data–masking–get–started–portal/.
[7]. A Globally Optimal k–Anonymity Method for the De–Identification of Health Data [Electronic resource] / [K. El Emam, F. Dankar, R. Issa a.o.]. – 2009. – Mode access: http://jamia.oxfordjournals.org/content/16/5/670.full.
[8]. Haldeman J. Compare IBM data masking solutions: InfoSphere Optim and DataStage [Electronic resource] / John Haldeman. – 2012. – Mode access: http://www.ibm.com/developerworks/data/library/techarticle/dm–1211maskingsolution/.

ЗАБЕЗПЕЧЕННЯ ЦІЛІСНОСТІ ЗОВНІШНІХ КЛЮЧІВ МАСКОВАНОЇ БАЗИ ДАНИХ

Стаття присвячена актуальній проблемі захисту інформаці у базах даних. Автори розглядають метод захисту даних шляхом маскування. Суть маскування даних полягає у безповоротній заміні у базі даних конфіденційної інформації (наприклад, даних, що ідентифікують конкретних людей) несекретними даними для відвертання доступу до неї неавторизованими користувачами. Як правило, конфіденційні дані замінюються схожими на реальні значення, щоб їх можна було використати в тестових системах з гарантією, що первинні дані не можуть бути отримані, витягнуті або відновленні. Маскування даних полягає в заміні вихідних конфіденційних даних (наприклад, даних, що ідентифікують конкретних людей) на інші, що приховують вихідні значення. Масковані дані повинні залишатися функціонально придатними для прикладної обробки (в основному, задач тестування, навчання тощо). У даній статті автори пропонують методику статичного маскування даних методом підстановки, що забезпечує маскування первинного ключа і цілісність зовнішніх ключів бази даних (БД). Дана методика реалізована у вигляді збереженої процедури для системи управління базами даних (СУБД) MS SQL Server.

Ключові слова: база даних, захист персональних даних, маскування даних, конфіденційні дані, інформаційна система.

INTEGRITY CONTROL OF MASKED DATABASE FOREIGN KEY

The paper is about the actual problem of information protection in databases. The authors examine the method of data protection by masking. The essence of data masking is irreversible replacement of confidential information in the database (such as data identifying specific people) with non–sensitive data to prevent access of unauthorized users. Normally, the confidential data is replaced with data similar to real values that they can be used in the test systems with the assurance that the original data can not be obtained, extracted or recovered. Data masking involve replacing the initial confidential data (such as identifying specific people data) to other, hiding the original values. The masked data must remain functionally suitable for application processing (mainly testing problems, tuition, etc.). In this paper, the authors propose a method of static data masking through substitution, providing masking the primary key and foreign keys integrity of the database. This method is implemented as a stored procedure for data base management system (DBMS) MS SQL Server.

Keywords: database, personal data protection, data masking, private data, information system.

Коломиец Михайло Володимирович, кандидат технічних наук, доцент Фізико-технічного інституту НТУУ «КПІ».
Коломыщев Михаил Владимирович, кандидат технических наук, доцент Физико-технического института НТУУ «КПИ».
Kolomytsev Myhailo, candidate of technical sciences, associate professor of Physico-Technical Institute of the NTUU "KPI".
E-mail:box144a@ukr.net

Носок Світлана Олександрівна, кандидат технічних наук, доцент Фізико-технічного інституту НТУУ «КПІ».
Носок Светлана Александровна, кандидат технических наук, доцент Физико-технического института НТУУ «КПИ».
Nosok Svitlana, candidate of technical sciences, associate professor of Physico-Technical Institute of the NTUU "KPI".
E-mail: svetlana@pti.kpi.net

Мазуренко Анастасія Євгенівна, студентка Фізико-технічного інституту НТУУ «КПІ».
Мазуренко Анастасия Евгениевна, студентка Физико-технического института НТУУ «КПИ».
Mazurenko Anastasia, student of the Physico-Technical Institute of the NTUU "KPI".
E-mail:ks0610@mail.ru