ISSN 2007-9737

La evaluación de la calidad de datos: una aproximación criptográfica

Jessica Yanes Pavón, Roberto Sepúlveda Lima, Humberto Díaz Pando

Universidad Tecnológica de La Habana "José Antonio Echeverría", CUJAE, Cuba

{jyanes, sepul, hdiazp}@ceis.cujae.edu.cu

Resumen. Los sistemas informáticos son un punto clave para la toma de decisiones en cualquier empresa que pretenda ser competitiva. Ello se debe a que estos sistemas se sustentan sobre la base de gestionar datos para la generación de información útil en la de toma de decisiones. En este contexto, se hace necesario que la información generada provenga de datos con la calidad adecuada, siendo la integridad, uno de los atributos más relevantes. Diversos han sido los enfoques y métodos estudiados en la bibliografía, los cuales persiguen como objetivos fundamentales identificar los escenarios donde pueden desencadenarse problemas en la calidad de los datos, así como definir métodos y mecanismos para garantizarla y medirla. Sin embargo, no se toma en cuenta en los trabajos analizados el escenario donde se vea afectada la calidad debido a amenazas de seguridad. El artículo se fundamenta en los criterios generales de la seguridad como atributo de calidad de todo sistema informático, particularizando en un enfoque de la seguridad de los datos como atributo de la calidad de estos. La principal contribución radica en la definición de un nuevo contexto de seguridad en el cual pueden surgir los problemas de calidad de datos. Se propone un método resistente a ataques para medir la calidad de los datos, basado en mecanismos criptográficos. Además, se realiza un análisis de la incidencia del método propuesto en los tiempos de respuesta del sistema. Los resultados obtenidos en la experimentación muestran que los aumentos de los tiempos no son significativos, por lo que no afectan de manera apreciable la disponibilidad de la aplicación objeto de estudio.

Palabras clave. Calidad de datos, evaluación de calidad, integridad de datos.

Evaluation of Data Quality: A Cryptographic Approach

Abstract. The computer systems are a key point for decision making in any company that intends to be competitive. This is because these systems are based on the management of data by generating useful information in the decision-making process. In this context, it is necessary that the information generated comes from data with the appropriate quality, being integrity one of the most relevant attributes. Diverse approaches and methods have been studied in the bibliography, which pursue as fundamental objectives to identify the scenarios where problems can be triggered in the quality of the data, as well as to define methods and mechanisms to guarantee and measure it. However, the scenario where quality is affected due to security threats is not taken into account in the analyzed works. The article is based on the general criteria of safety as an attribute of quality of any computer system, particularizing in a data security approach as an attribute of the quality of these. The main contribution lies in the definition of a new security context in which data quality problems can arise. A resistant method to attacks is proposed to measure the quality of the data, based on cryptographic mechanisms. In addition, an analysis is made of the incidence of the proposed method in the response times of the system. The results obtained in the experimentation show that the increases of the times are not significant, so they do not appreciably affect the system availability.

Keywords. Data quality, quality assessment, data integrity.

1. Introducción

Los sistemas de información han demostrado ser un fuerte punto de apoyo para el procesamiento de la información en todos los sectores y esferas de la vida actual. Estos agilizan el proceso de gestión del negocio, ya que permiten organizar la forma en que es procesada la información [1].

Como parte de su evolución lógica, los volúmenes de datos que estos sistemas gestionan van en aumento junto a una creciente dependencia de las empresas para lograr elevar su confiabilidad y competitividad, y una cierta incertidumbre acerca de la posible obsolescencia de estos.

Con el decursar de los años la información almacenada establece la historia y el presente del negocio, constituyendo el pilar fundamental sobre el que se soporta la toma de decisiones futuras de la organización [2, 3]. Para tomar una buena decisión es de vital importancia contar con sistemas que se ajusten a las características de los procesos y que sean confiables [2]. Para que un sistema sea confiable debe garantizar, entre otros aspectos, que sus datos tengan la calidad requerida por los usuarios de forma tal que sea posible generar información útil y confiable a partir de ellos.

A partir de lo anterior se hace necesario monitorear la calidad de los datos almacenados, de modo que puedan ser identificadas las anomalías y establecer un proceso de corrección sobre las mismas, evitando así la toma de decisiones sobre la base de información incorrecta. En el año 2003 Dasu et al. [4] plantearon que es común que entre un 60% y 90% de los datos almacenados en una base de datos tengan problemas de calidad. En el entorno de las bases de datos la tarea de evaluar[1] la calidad de los datos sin la ayuda de un método o herramienta especializada puede resultar un tanto tediosa, debido a la gran cantidad de datos almacenados, requiriendo mucho tiempo y esfuerzo para su análisis.

Ante la notable necesidad de garantizar la calidad de la información se desarrolla una nueva área de investigación centrada en el estudio de la calidad de datos. En la ISO/IEC 25012:2008 [5], la calidad de datos queda definida como el grado en que los datos satisfacen las necesidades de los usuarios. Está compuesta por distintos aspectos conocidos como dimensiones de calidad que permiten estudiarla o analizarla en forma más detallada de la calidad de datos. Estas dimensiones ofrecen una forma de medir y gestionar la calidad de los datos [6]. La calidad de datos puede ser evaluada mediante el uso de métricas, a través de las cuales se define la forma en la que cada dimensión es medida [7].

Existen diversos factores por los cuales una aplicación puede presentar datos con mala calidad [8]. El área de investigación que estudia la temática se centra en la entrada de los datos y la procedencia de los mismos en caso de que se tengan varias fuentes de información. Sin embargo, no son estas las únicas formas en la cual los datos se pueden “ensuciar” o perder sus parámetros de calidad. En ninguno de los estudios se toma en cuenta la posibilidad de que los datos puedan ser “ensuciados” desde un contexto donde no se garantiza la seguridad de los datos.

Existe un escenario desde el cual la calidad de los datos puede ser afectada, y es desde el punto de vista de la seguridad. Los datos almacenados en una base de datos o cualquier otro almacén de datos, pueden estar expuestos a ser modificados de manera impropia por personal no autorizado y con altos privilegios. Como consecuencia de este fenómeno, la calidad de los datos almacenados puede ser comprometida y, con ello, la información utilizada en la toma de decisiones. Por todo lo anterior, es posible afirmar la existencia de una relación entre la seguridad y la calidad de los datos, lo cual sugiere el empleo de técnicas criptográficas en aras de evaluar la calidad de éstos.

Diversos han sido los métodos estudiados enfocados a medir y garantizar la calidad de los datos [7, 9-15]. Cada uno de éstos permite identificar la ocurrencia de los problemas de calidad de datos en cada uno de los contextos estudiados anteriormente. Sin embargo, ninguno contempla el escenario en el que los datos se ensucian por violación de la integridad en el contexto de seguridad mencionado anteriormente. Esto trae consigo que los métodos desarrollados hasta la actualidad también sean vulnerables a este contexto. Al igual que el dato es vulnerado, el medio mediante el cual se verifica su calidad pudiera estar siendo modificado por el atacante.

[1] En la bibliografía consultada se utilizan indistintamente los términos evaluar y medir.

Como consecuencia de este fenómeno se estaría haciendo uso de datos erróneos sin ser detectado por las aplicaciones que los consumen.

A partir de lo anterior la principal contribución de este artículo radica en un método resistente a ataques para medir la calidad de los datos en las dimensiones exactitud, completitud y consistencia [7] de la calidad de los datos, fundamentado en mecanismos criptográficos. Los resultados obtenidos en la experimentación muestran la eficacia del método detectando modificaciones impropias en los datos. Se evidencia un ligero aumento en los tiempos de respuesta por parte del sistema bajo experimentación, los cuales no afectan notablemente la disponibilidad del mismo.

El resto del artículo queda estructurado de la siguiente manera: en la sección 2 se presenta una aproximación a los fundamentos básicos acerca de la calidad de datos, así como un estudio de los contextos en los cuales los datos pierden calidad. En la sección 3 se describe la técnica propuesta para evaluar la calidad que exhiben los datos. En la sección 4 se realiza la validación de la propuesta mediante un caso ejemplo. Se realiza la discusión de los resultados obtenidos a partir de aplicar la propuesta y por último se presentan las conclusiones y trabajos futuros.

2. Trabajos relacionados

Varios autores han establecido una definición para la calidad de datos [5, 8, 16]. Sin embargo, todos concuerdan en que la calidad de datos está compuesta por distintas características, comúnmente conocidas como dimensiones de calidad, que reflejan un aspecto particular.

En los artículos consultados [17-20], concurre un núcleo de dimensiones que es compartido por la mayoría de los autores. La ISO/IEC 25012:2008 [5] define un modelo de calidad de datos conformado por quince dimensiones desde dos puntos de vista: (1) inherente, el cual se refiere al dato en sí y a la correspondencia del mismo con la información del mundo real y (2) dependiente del sistema. El segundo punto de vista es dependiente de la plataforma tecnológica en la cual es empleada la información. Cada una de las dimensiones corresponde a uno de los dos puntos de vista anteriormente expuestos, aunque existe un conjunto de dimensiones que tienen correspondencia con ambos tal y como se muestra en la tabla 1. Las dimensiones objetivo en el presente artículo son: exactitud, completitud y consistencia.

Tabla 1. Dimensiones de la calidad de datos según ISO 25012:2008

Características	Puntos de vista de la CD	
	Inherente	**Dependiente del sistema**
Exactitud	X	
Completitud	X	
Consistencia	X	
Credibilidad	X	
Actualidad	X	
Accesibilidad	X	X
Conformidad	X	X
Confidencialidad	X	X
Eficiencia	X	X
Precisión	X	X
Trazabilidad	X	X
Comprensibilidad	X	X
Disponibilidad		X
Portabilidad		X
Recuperabilidad		X

Exactitud: indica la precisión de los datos, que tan precisos y válidos son [5].

Completitud: esta dimensión se refiere a que el sistema de información contiene todos los datos que son de interés para los procesos del negocio [5].

Consistencia: esta dimensión está relacionada a la coherencia existente entre un dato y el resto [5].

Estas son las dimensiones de calidad de datos que más relación tienen con el atributo de la integridad considerado por la seguridad de

Tabla 2. Problemas de calidad de datos según [8]

Problemas de calidad de datos	Atributo- Fila			SR	MR	MDS
	SAST	SMT	MAST			
Valor ausente	x					
Violación de sintaxis	x					
Valor incorrecto	x					
Violación de domino	x					
Subcadena invalida	x					
Errores ortográficos	x					
Valor impreciso	x					
Violación de restricciones de dominio	x	x	x	x	x	x
Violación de restricción de unicidad		x				
Existencia de sinónimos		x				
Fila semivacía			x			
Violación de dependencias funcionales			x			
Filas aproximadamente duplicadas				x		
Filas duplicadas inconsistentes				x		
Violación de integridad referencial					x	
Referencia incorrecta					x	
Inconsistencia en la sintaxis					x	x
Circularidad entre filas					x	
Inconsistencia en las unidades de medida						x
Heterogeneidad en la representación						x
Existencia de homónimos						x

datos, la que puede ser definida como la protección de la información contra la modificación accidental o intencional no autorizada que puede afectar la validez de los datos [29].

Los problemas de calidad de datos son conocidos como anomalías, errores o incluso suciedad [8]. En cuanto a los problemas de calidad de datos, son múltiples las investigaciones realizadas, en las cuales han sido definidos cada uno de los problemas identificados [8, 21, 22, 23].

Según [8] los problemas de calidad de datos relacionados a los atributos y las filas pueden ser agrupados en cinco niveles fundamentales de granularidad: único atributo en una única fila (SAST), único atributo en múltiples filas (SAMT, una columna), múltiples atributos en una única fila (MAST, una fila), múltiples relaciones (MR) y múltiples fuentes de datos (MDS). La agrupación se realiza siguiendo un enfoque ascendente desde un nivel más bajo (atributo/fila) hasta un nivel más alto (múltiples fuentes de datos). En la tabla 2 se ilustra el dominio común de problemas de calidad de datos identificados en la bibliografía consultada.

Cada uno de los problemas identificados son agrupados en los distintos niveles de granularidad en los cuales pueden surgir. Nótese que, para el caso de la violación de restricción del dominio es un problema que puede surgir tanto en los niveles más bajos y específicos de los

datos como en los niveles más generales de granularidad.

De igual modo cada uno de los niveles de granularidad antes identificados constituye una vía desde la cual pueden ser solucionados dichos problemas.

Los problemas de calidad de datos mayoritariamente se materializan en tres contextos diferentes, un primer contexto cuando existen anomalías en una única fuente de datos, el segundo está dado cuando datos no estructurados son migrados a una fuente de datos estructurada y por último cuando se realizan procesos de integración en una única fuente datos de información proveniente de diversas fuentes [8]. Cada uno de estos contextos constituyen amenazas que desencadenan en problemas de calidad en los diferentes niveles de granularidad mencionados anteriormente, los que producen afectaciones en las dimensiones de calidad de datos.

Para conocer el grado de calidad con que cuentan los datos de una organización es imprescindible llevar a cabo un proceso de evaluación de la calidad. La evaluación es el proceso mediante el cual se identifican datos relacionados a un elemento específico que permiten establecer criterios para valorar hasta qué punto estos elementos cumplen con los fines y objetivos establecidos.

En la literatura [7,9-15,24-26] es posible encontrar varias técnicas orientadas todas a los procesos de medición y mejoramiento de la calidad de los datos. A continuación, se hace mención de las técnicas más utilizadas en el proceso de evaluación de la calidad de los datos.

Uso de métricas

Son utilizadas para medir las distintas dimensiones de la calidad de datos [7]. Una métrica define la forma en la que un factor de calidad es medido [9]. Las métricas de evaluación son heurísticas diseñadas para ajustarse a una situación de evaluación específica [24]. Son diversos los artículos que proponen procesos de evaluación de la calidad basados en métricas [7, 9-12]. Un factor de calidad puede ser medido por un conjunto de métricas, así como una misma métrica puede ser utilizada para medir diferentes factores de calidad.

Las métricas pueden estar definidas por funciones que realicen comparación entre datos, cálculo de la distancia entre dos valores [13], etc. Los resultados de la aplicación de estas funciones proveen un valor numérico que determinan el rango de ocurrencia de determinados problemas en la dimensión medida [25]. Una métrica puede estar definida por un valor lógico que indica si un dato es sintácticamente correcto o no, tomando valor 0 para el caso correcto y 1 en caso de que no sea correcto el dato, por ejemplo.

El empleo de métricas es una de las técnicas más populares en los procesos de evaluación de la calidad de datos. No obstante, dicha técnica es vulnerable a ataques por parte de personal interno a la organización.

Uso de metadatos

Los metadatos son archivos de información que recogen las características básicas de algún dato o recurso. En este caso no solo es almacenado el dato, sino que también junto al dato es almacenado su metadato, en el que se contiene la descripción detallada del dato, su composición, así como su origen, entre otras características relevantes del dato. Para el caso en el que un dato es modificado incurriendo en una violación de la integridad no existirá correspondencia entre el dato modificado y el metadato almacenado [26].

La debilidad de esta técnica radica en que, si el metadato no se encuentra protegido, el atacante además de modificar el dato pudiera estar modificando a su vez el metadato que lo describe, no detectándose el ataque. Ante la ocurrencia de un ataque de este tipo se obtendría un resultado erróneo del proceso de evaluación.

Uso de ontologías

En la mayoría de los casos, las ontologías son utilizadas para describir el conocimiento de un dominio determinado [14]. Las ontologías permiten el desarrollo sistemático de métodos automatizados válidos y confiables para evaluar los problemas de calidad de datos [15]. Los beneficios de utilizar una ontología para describir la calidad de los datos son que una ontología está escrita en un lenguaje formal, es capaz de representar la semántica, proporciona un

vocabulario compartido para discutir la calidad de los datos y es suficientemente riguroso para ser utilizado directamente en algoritmos y computadoras [10]. Sin embargo, el uso de esta técnica también es susceptible a ataques similares al expuesto con anterioridad. Hasta el momento no se evidencia en los artículos revisados el uso de ningún mecanismo que proteja la ontología en este sentido.

De acuerdo a la situación actual de los ataques a los sistemas informáticos [27, 28], no se aprecia en la literatura un estudio realizado desde un contexto de seguridad, en el cual pueden surgir los problemas de calidad de datos, quedando totalmente comprometida como producto de una amenaza de seguridad. De igual modo pueden ser alteradas las técnicas de evaluación anteriormente expuestas.

Este contexto de seguridad está definido por un modelo de amenazas en el cual se identifica como requisito de seguridad clave la integridad, son identificados como recursos vulnerables los datos relevantes que presenta una organización, y las técnicas de evaluación de los niveles de calidad con que cuentan los mismos. La amenaza está constituida por la ocurrencia de acciones deliberadas, por parte de personal perteneciente a la organización y con determinados privilegios, dirigidas a modificar los datos. De ocurrir estas acciones, se violaría la integridad de los datos, comprometiendo así la calidad de los mismos, teniendo un elevado impacto negativo en los procesos que ocurren en la organización.

La situación anterior evidencia la existencia de una relación entre la integridad de datos desde el entorno de la seguridad y la calidad de los mismos. Esta relación permite el empleo de mecanismos de integridad para garantizar los atributos de exactitud, completitud y consistencia de la calidad de datos. Al emplear estos mecanismos de seguridad se estaría mitigando el riesgo.

Ante la situación anterior se realizó un estudio acerca de los mecanismos existentes para garantizar la integridad de los datos más allá de los que incorporan los sistemas gestores de bases de datos y sistemas operativos, con la finalidad de proveer un método de medición de calidad de datos resistente en el contexto de seguridad. En la figura 1 se muestran los mecanismos estudiados clasificados según el estado final del mensaje. A continuación, se describirán los mecanismos empleados en este artículo. Se distinguen los que manejan texto claro y los que manejan texto cifrado.

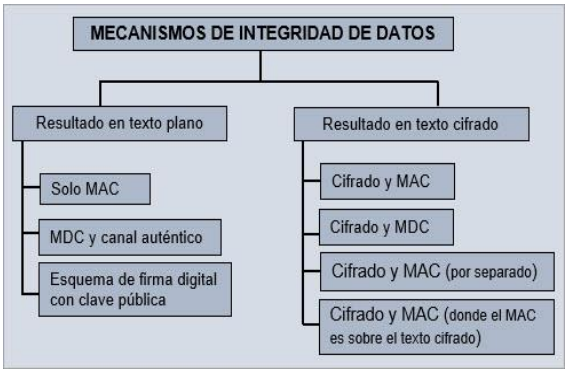

Fig. 1. Mecanismos de control de integridad según estado final del mensaje

Solo MAC[2]

El emisor del mensaje x calcula un MAC $h_k(x)$ sobre el mensaje usando una clave secreta k, donde esta además es compartida con el receptor. Luego envía el mensaje y el valor MAC $(x| \ |h_k(x))$[3], donde ambos están divididos por algún identificador. Una vez que llega al receptor, este separa el mensaje y el valor MAC recibido, calcula el MAC del mensaje de forma independiente usando la clave compartida con el emisor y compara el MAC calculado con el recibido. Si ambos valores son iguales entonces el receptor tiene la garantía de que el mensaje es auténtico e íntegro. Por tanto, el mensaje fue generado por una parte que conoce la clave compartida y el mensaje no fue alterado [29, 30].

Cifrado y MAC

Esta técnica puede ser empleada en escenarios donde tanto la confidencialidad como la integridad son requeridas. El emisor del mensaje x calcula un valor MAC $\mathrm{h_k(x)}$ sobre el

[2] Del inglés: Códigos de Autenticación de Mensajes.

[3] Concatenando el mensaje con la codificación correspondiente.

mensaje usando una clave secreta k, donde esta además, es compartida con el receptor. Luego cifra el paquete usando un algoritmo de cifrado simétrico E y una clave compartida k' . El texto cifrado es producido por la siguiente expresión $C = E_{k'}(x \mid \ |h_k(x))$. En determinado escenario $k' = k$, lo cual no es recomendable por cuestiones de seguridad. El paquete cifrado C es enviado por un canal inseguro [29, 31, 32].

Cifrado y MDC4

El emisor del mensaje x, calcula un valor MDC $H = h(x)$ sobre el mensaje x. Adiciona el valor MDC al mensaje x, y cifra el paquete usando un algoritmo de cifrado simétrico E con una clave compartida k. La siguiente expresión produciría el texto cifrado $C = E_k(x| \ |h(x))$. El paquete cifrado C es enviado por un canal inseguro. En el otro extremo, el receptor descifra el paquete C, separando el mensaje x' y el valor MDC H' recibidos (están separados por algún identificador).

El receptor calcula de forma independiente el valor MDC del mensaje recibido x' y lo compara con el valor MDC recibido H'. Si ambos valores son iguales entonces el receptor tiene la garantía que el mensaje es auténtico e íntegro. La intención es que el cifrado proteja el valor de compresión unidireccional añadido (hash) y que sea imposible que un atacante sin la clave de cifrado altere el mensaje sin interrumpir la correspondencia entre el valor obtenido del cifrado y el MDC recuperado [29].

3. Propuesta

El objetivo de esta sección es describir los elementos fundamentales del método propuesto. El método propuesto permitirá medir la calidad de datos para las dimensiones exactitud, completitud y consistencia. Este método permitirá identificar la pérdida de la calidad de los datos en estas dimensiones en cualquiera de los cuatro contextos descritos previamente.

[4] Del inglés: Código de Detección de Mensajes

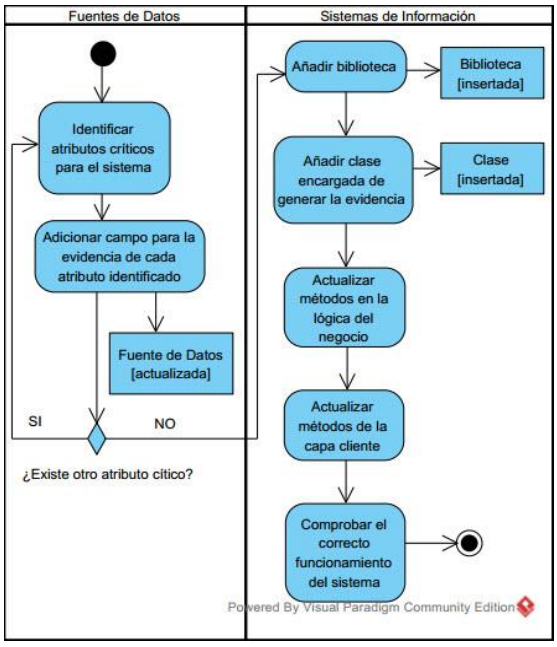

Fig. 2. Flujo de tareas para integrar el método de medición de calidad de datos a un sistema

El método de medición queda definido por la siguiente expresión (1):

$$MC \equiv (MD, F, \Gamma(\mathrm{MD}, \mathrm{F})), \quad (1)$$

donde MD es el conjunto de posibles modelos de datos asociados a los problemas de calidad de datos y se define por: $MD = \{SAST, SAMT, MAST, SR, MR, MDS\}$

A su vez F, representa el dominio de funciones para generar y verificar la evidencia y queda definida en la expresión (2):

$$F \equiv (f_G(d), f_v(x, d)). \quad (2)$$

donde d representa el dato al cual se medirá la calidad, mientras que F está compuesta por funciones criptográficas seguras como son las funciones de HASH, MAC, MDC, etc. La función $f_G(d)$ es la encargada de generar el valor de evidencia de la forma $x = f_G(d)$ a partir del dato d; mientras que la función $f_v(x, d)$ se define como (3):

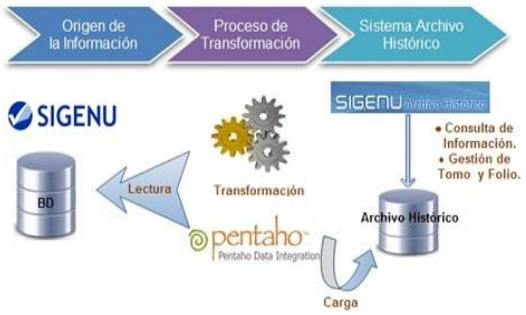

Fig. 3. Migración de información hacia base de datos de Archivo Histórico

$$f_v(x,d) = \begin{cases} 0, & x \neq f_G(d) \\ 1, & x = f_G(d) \end{cases}. \quad (1)$$

Por último, $\Gamma(MD,F)$, representan las transformaciones a ejecutar en el esquema de datos, en dependencia del modelo de datos y de las funciones utilizadas para incorporar el método a la aplicación. La integración del método propuesto a un sistema implica variaciones en los tiempos de respuesta del sistema, así como en el crecimiento de la base de datos, aspectos que pueden afectar la disponibilidad.

En la figura 2 es ilustrado el flujo de tareas a realizar para incorporar el método a un sistema. Primeramente, deberán ser identificadas las tablas que contienen datos vulnerables para el sistema. Seguidamente para cada uno de los datos identificados se agrega un campo que almacenará el valor de redundancia correspondiente al dato.

En el sistema en cuestión será necesario adicionar una biblioteca que provea los métodos criptográficos necesarios. Luego tendrán que ser modificados los métodos mediante los cuales la información almacenada en la fila es insertada, modificada y recuperada, a fin de incorporar un nuevo atributo redundante. Finalmente se comprueba el correcto funcionamiento del sistema luego de incluida la modificación propuesta.

4. Experimentos

La experimentación persigue dos objetivos fundamentales: (1) evaluar la eficacia del método detectando modificaciones impropias en los datos, (2) realizar un análisis de la incidencia del método propuesto en los tiempos de respuesta del sistema, así como en el crecimiento del volumen de la base de datos.

Escenario de experimentación

El escenario de experimentación lo conforma el Sistema de Gestión de la Nueva Universidad en lo adelante SIGENU, sistema mediante el cual se gestiona la información en las universidades cubanas. El proceso de experimentación se llevó a cabo específicamente en el Módulo de Archivo Histórico del sistema SIGENU. Dicho módulo se especializa en la gestión de la información docente de los egresados que cursaron estudios en la institución. Una de las principales funciones del sistema es hacer persistir la información en el tiempo de manera tal que pueda ser consultada en tiempos futuros. Se almacena la información tanto de los graduados como de los estudiantes que causaron baja definitiva del centro de estudios [33].

Existen dos procesos fundamentales mediante los cuales el sistema se nutre de información. Un primer proceso en el cual es migrada toda la información relacionada con los egresados de la institución desde el Módulo de Secretaría del sistema SIGENU, como se aprecia en la figura 3.

El segundo proceso está determinado por la gestión de la información. Se introduce en el sistema la información de aquellos estudiantes que no se encuentran en el Módulo de Secretaria del sistema de gestión docente y su información se encuentra solamente registrada en expedientes en formato duro.

En esta base de datos, fue identificado como dato sensible la calificación de los estudiantes en cada una de las asignaturas cursadas durante su trayecto por el centro de educación superior. En este caso para cada una de las notas a proteger se generó una evidencia, que permite identificar la ocurrencia de una modificación sobre el valor original. La implementación del método es realizada utilizando un modelo de datos único atributo de una única fila (SAST), debido a que hasta el momento solo se identificaron las notas como dato sensible.

La evidencia está compuesta por el valor del atributo junto con su valor de redundancia.

La evidencia es generada mediante el empleo de mecanismos criptográficos. En la realización de las pruebas fueron empleados solo MAC, Cifrado y MAC y Cifrado y MDC como mecanismos de integridad. La clave necesaria por los algoritmos criptográficos es conformada con elementos del identificador del estudiante y elementos del identificador de la asignatura matriculada. Para la puesta en práctica del método propuesto fue empleada la biblioteca *bouncycastle*[5], la cual proporciona un conjunto de métodos criptográficos.

Se identifican cuatro funciones en el sistema las cuales se encargan de la gestión de las notas de los estudiantes en las asignaturas matriculadas:

getMatriculatedSubjectByStudent en lo adelante F1: función que permite listar las asignaturas matriculadas de un estudiante determinado.

addSubjectByStudent en lo adelante F2: esta es la función a través de la cual le es matriculada una asignatura a un egresado.

updateMatriculatedSubject en lo adelante F3: función a emplear cuando se quieren modificar los datos de una asignatura que ya se le ha matriculado a un estudiante.

addSubjectByStudentLote en lo adelante F4: esta función permite que más de una asignatura sea matriculada a más de un estudiante en el mismo momento.

4.1. Eficacia del método

Una vez integrado el método propuesto al sistema se procedió a registrar las notas de las asignaturas matriculadas de un estudiante. Para cada nota de las asignaturas fue generada la evidencia correspondiente.

Para evaluar la eficacia del método ante ataques se modificaron las notas de parte de las asignaturas matriculadas accediendo directamente a la base de datos.

Cuando es consultado nuevamente el listado de asignaturas del estudiante aparecen identificados en color rojo los registros correspondientes a las asignaturas modificadas. De igual modo ocurre si en lugar de cambiar el valor de la nota cambiamos el valor de evidencia correspondiente.

Con la realización de este experimento se valida la correcta generación de la evidencia, así como la detección de modificaciones impropias sobre los datos. De este modo se valida que el método es resistente al contesto de seguridad expuesto en secciones anteriores.

Tabla 3. Valores p obtenidos de la prueba estadística

	F1	F2	F3	F4
Solo MAC	0.370	0.272	0.064	0.155
Cifrado y MAC	0.003	0.080	0.020	0.014
Cifrado y MDC	0.021	0.168	0.408	0.166

4.2. Impacto en el rendimiento

Para la evaluación del impacto de la solución propuesta en el rendimiento del sistema se llevó a cabo un proceso de experimentación con el objetivo de realizar un análisis de la incidencia del método propuesto en los tiempos de respuesta del sistema. Fueron realizadas 10 ejecuciones para cada uno de los procesos que gestionan los datos vulnerables antes y después de añadir cada uno de los métodos de integridad analizados. Los experimentos fueron realizados sobre la base de datos del Sistema de Archivo Histórico del SIGENU. Actualmente la tabla donde fue identificado el dato sensible cuenta con una población total de 483038 registros.

Para comprobar si la inclusión de los métodos estudiados produce cambios en los tiempos de ejecución del sistema se llevó a cabo una prueba de hipótesis para cada uno de estos. A continuación, se describe la prueba realizada.

Las hipótesis evaluadas fueron:

H_0: T_s-T_m =0

H_1: T_s-T_m ≠0

Donde:

H_0: Hipótesis nula.

[5] Biblioteca de Java que complementa la extensión criptográfica Java (JCE) predeterminada.

H_1: Hipótesis alternativa.

α: Máximo nivel de riesgo aceptable para rechazar una hipótesis nula verdadera. En este 0.5.

T_s: Tiempo de respuesta del sistema sin el método incorporado.

T_m: Tiempo de respuesta del sistema luego de incorporar el método.

Para realizar la prueba de las hipótesis fue utilizada la prueba no paramétrica de Mann-Whitney. Fue seleccionada esta prueba debido a que no existe una evidencia que de los datos sigan una distribución normal. La realización de esta prueba arroja un resultado p, el cual indica la probabilidad de obtener la mediana de la muestra si H_0 es verdadera.

La prueba fue realizada mediante el uso del software estadístico Minitab. La realización del experimento arrojó los datos mostrados en la tabla 3.

Como resultado del análisis se puede afirmar:

- El desarrollo del método propuesto mediante el empleo de la técnica de solo MAC no implica un aumento significativo en los tiempos de respuesta del sistema. Como se puede apreciar en la tabla 3 los valores obtenidos son mayores que el nivel de riesgo aceptado por lo cual no puede ser rechazada la hipótesis nula.
- El empleo de la técnica Cifrado y MAC implica un aumento significativo de los tiempos de respuesta cuando son ejecutadas F1, F3, F4. Para el caso de la función F2 el aumento de los tiempos no es significativo.
- La puesta en práctica de la solución propuesta haciendo uso de Cifrado con MDC representa un aumento significativo en la ejecución de la función F1. En el caso de las funciones restantes el aumento implicado no es significativo.

4.3. Impacto en el tamaño de la base de datos

El volumen de información con que cuentan las bases de datos de los sistemas es uno de los aspectos fundamentales que valoran las organizaciones ante la implantación de nuevas soluciones. Es por lo anterior que se realiza un análisis de la redundancia que se añade a la base de datos del sistema una vez que es incorporada la propuesta de solución. A continuación, se detalla el comportamiento del crecimiento de la base de datos para cada uno de los métodos analizados. El incremento del tamaño está dado por la expresión 4:

Tabla 4. Incremento de la fuente de datos

Mecanismo de integridad	CE	CA	TR	TA
Solo MAC	1400	69	520 bits	5 MB[6]
Cifrado y MAC			520 bits	5 MB
Cifrado y MDC			392 bits	4 MB

$$TA = CE \cdot CA \cdot TR, \tag{4}$$

donde:

TA: representa el tamaño agregado.

CE: representa la cantidad de estudiantes graduados.

CA: está dada por la cantidad de asignaturas cursadas por un estudiante graduado.

TR: representa el tamaño del valor de redundancia.

Anualmente el promedio de graduados es de 1400 estudiantes. Para cada uno de estos son almacenadas las asignaturas cursadas durante su recorrido por la institución. Los estudiantes graduados poseen un promedio de 69 asignaturas cursadas. Una vez implantada la propuesta de solución cada asignatura contará con un valor de redundancia asociado a la evaluación de la misma.

En la tabla 4 se muestran los resultados obtenidos luego de aplicar la expresión 4.

El empleo de los mecanismos de integridad MAC o CMAC implican un incremento de 5MB. Por otra parte, al emplear CMDC se obtiene un incremento de la base de datos de 4MB.

[6] Megabyte (MB) una unidad de medida de almacenamiento.

5. Conclusiones

El buen desempeño y crecimiento de las organizaciones se propicia mediante la correcta toma de decisiones por parte de las mismas. Para ello el elemento clave lo constituyen los datos sobre los cuales se sustentan las decisiones a tomar por los directivos de la organización. Siendo estos datos la clave del éxito de las organizaciones es de vital importancia garantizar que cuentan con el nivel de calidad adecuado.

En el artículo fue analizado un nuevo contexto desde el cual la calidad de los datos se puede ver afectada. El nuevo contexto estudia la calidad desde un punto de vista de la seguridad en el cual la realización de modificaciones impropias ensuciaría los datos almacenados.

Se expuso un método que permite evaluar la calidad desde este contexto de seguridad. En todos los escenarios de experimentación del método con el empleo de cada uno de los mecanismos estudiados quedo validada la resistencia del mismo al contesto de seguridad, así como su eficacia.

A partir de los resultados obtenidos se puede concluir que el aumento de los tiempos es significativo para el caso de las funciones en las cuales son manipuladas más de una asignatura, como la matricula en lotes y el listado de asignaturas.

Referencias

1. **Ramírez, J. & Vega, O. (2015).** Sistemas de información gerencial e innovación para el desarrollo de las organizaciones. *Telematique 14.*
2. **Sánchez-Godínez, E. & Súñiga-Segura, L. (2011).** La importancia de contar con información precisa, confiable y oportuna en las bases de datos. *Revista Nacional de Administración,* Vol. 2, No. 2, pp. 145–154. DOI: 10.22458/rna.v2i2.377.
3. **Subiela-Durá, S. (2011).** *Sistemas de Información BI: Estado actual y herramientas de software libre.*
4. **DASU, T.E.A. (2003).** Data quality through knowledge engineering. *Ninth International Conference on Knowledge Discovery and Data Mining ACM SIGKDD.*
5. **ISO/IEC (2008).** *Software Engineering — Software product Quality Requirements and Evaluation (SQuaRE).* Data quality model.
6. **Sidi, F., Panahy, P.H.S., Suriani, A.L., Jabar, M. A., & Ibrahi, H. (2012).** *Data Quality: A Survey of Data Quality Dimensions*. IEEE. DOI:10.1109/ InfRKM.2012.6204995.
7. **Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2012).** *Quality Assessment for Linked Open Data: A Survey.*
8. **Oliveira, P., Rodrigues, F., & Henriques, P. (2005).** A Formal Definition of Data Quality Problems. *International Conference on Information Quality.*
9. **Marotta, A., Vallespir, D., & Valverde, C. (2012).** Análisis de la Calidad de Datos en Experimentos en Ingeniería de Software.
10. **Pipino, L., Lee, Y., & Wang, R. (2002).** Data Quality Assessment. *Communication (ACM),* Vol. 45, No. 4. pp. 211–218.
11. **Kaiser, M., Klier, M., & Heinrich, B. (2007).** How to Measure Data Quality? - A Metric-Based Approach. *International Conference on Information Systems (ICIS),* pp. 1–15.
12. **Lin, M. & Hua. Z. (2008).** A Method for Measuring data quality in Data Integration. *International Seminar on Future Information Technology and Management Engineering.* DOI:10.1109/FITME. 2008.146.
13. **Otto, B., Hüner, K., & Österle, H. (2009).** Identification of Business Oriented Data Quality Metrics. *Presented at the ICIQ,* pp. 122–134.
14. **Barchini, G., Alvarez, M., Palliotto, D., & Fortea, G. (2009).** Evaluación de la calidad de los sistemas de información basados en ontologías. *IX Congreso ISKO-España.*
15. **Liawabc, S.T., Rahimi, A., Ray, P., Taggart, J.R., Dennis, S., Lusignan, S., Jalaludin, B., Yeo, A. E.T., & Talaei-Khoei, Amir. (2012).** Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. *International Journal of Medical Informatics,* Vol. 82, No. 1, pp. 10–24. DOI: 10.1016/j.ijmedinf.2012.10.001.
16. **Guerra-García, C., Menéndez-Domínguez, Caballero, I, & Montaño, O. (2015).** Montaño Selecting web functionalities versus data quality dimensions: A first approach. *5th International Symposium on Data-driven Process Discovery and Analysis SIMPDA,* pp. 131–136.
17. **Lee, Y.W., Strong, D.M., Kahn, B.K., & Wangd, R.Y. (2002)**. AIMQ: a methodology for information quality assessment. *Elsevier Science,* Vol. 40, No. 2, pp. 133–146. DOI: 10.1016/S0378-7206(02) 00043-5.

18. **Neely, M.P. (2005).** The product approach to data quality and fitness for use: A framework for analysis. *10th International Conference on Information Quality MIT,* pp. 52–66.
19. **Lee, Y., Strong, D., & Wang, R. (1997).** Data quality in context. *Communications of the ACM,* Vol. 40, No. 5, pp. 103–110.
20. **Wang, R., Reddy, M., & Kon, H. (1995).** Toward quality data: an attribute-based approach. *Elsevier Science. Decision Support Systems,* Vol. 13, No. 3, pp. 349–372. DOI: 10.1016/0167-9236 (93)E0050-N
21. **Mouzhi, G. & Helfert, M. (2007).** A Review of Information Quality Research. *12th International Conference on Information Quality*, pp. 76–91.
22. **Oliveira, P., Rodrigues, F., Henriques, P., & Galhardas, H. (2005).** A Taxonomy of Data Quality Problems. *2nd Int. Workshop on Data and Information Quality,* pp. 219–233.
23. **Silva-Costa, T., Marques, B., & Freitas, A. (2010).** Data quality problems in administrative inpatient databases. *5th Iberian Conference on Information Systems and Technologies.*
24. **Bizer, C. (2007).** Quality-Driven Information Filtering in the Context of Web-Based Information Systems.
25. **Loshin, D. (2006).** Monitoring Data Quality Performance Using Data Quality Metrics.
26. **Verbo, E., Caballero, I., Perez, R., Calero, C., & Plattoni, M. (2008).** Una Metodología Basada en ISO/IEC 15939 para la Elaboración de Planes de Medición de Calidad de Datos. *13th Conference on Software Engineering and Databases, (JISBD´08),* pp. 253–264.
27. **Centeno-Ureña, F. J. (2015).** Ciberataques, la mayor amenaza actual. *Español de Estudios Estratégicos,* pp. 1–18.
28. **Sanchez-Madero, G. (2012).** La ciberguerra: los casos de Stuxnet y Anonymous. *Nueva Época,* pp. 124–133.
29. **Menezes, A., Oorschot, P.V., & Vanstone, S. (1996).** Handbook of Applied Cryptography.
30. **Preneel, B. (1988).** *The State of Cryptographic Hash Functions.* Lectures on Data Security, Modern Cryptology in Theory and Practice, Summer School, Aarhus, Denmark.
31. **Schneier, B. (1996).** *Applied Cryptography*. Protocols, Algorthms, and Source Code in C (cloth).
32. **Mao, W. (2003).** *Modern Cryptography: Theory and Practice.* Prentice Hall PTR.
33. **Yanes–Pavón, J. (2013).** *Versión web del Sistema de Gestión de la Nueva Universidad.* Módulo de Archivo Histórico, pp. 131.

Article received on 17/02/2018; accepted on 04/05/2018.
Corresponding author Jessica Yanes Pavón.