Ingeniería Industrial/ISSN 1815-5936/ Vol XL /No. 1/enero-abril /2019/ p. 37-47

ARTÍCULO ORIGINAL
INFORMÁTICA EMPRESARIAL

Evaluación de la calidad de datos mediante código de autenticación de mensajes

Evaluation of data quality through messages authentication code

Jessica Yanes-Pavón, Humberto Díaz-Pando

Universidad Tecnológica de La Habana, José Antonio Echeverría. La Habana, Cuba.
Correo eletrónico: jyanes@ceis.cujae.edu.cu, hdiazp@ceis.cujae.edu.cu

Recibido: 19 de febrero del 2018
Aprobado: 26 de octubre del 2018

RESUMEN
Los procesos de toma de decisiones brindan información de vital importancia para cualquier organización que pretenda ser competitiva. Los sistemas informáticos son fundamentales para un buen proceso de toma de decisiones. Es imprescindible que la información que sustenta las decisiones proceda de datos que cuenten con la calidad de datos adecuada, identificándose la integridad como uno de los atributos más importantes. Se sigue el principio de la seguridad cómo un atributo de calidad de los sistemas, particularizando en la seguridad de los datos y la calidad de los mismos. La principal contribución radica en el uso de código de autenticación de mensajes para la evaluación de calidad que exhiben los datos almacenados en una fuente de datos vulnerable. Los resultados obtenidos en la experimentación revelan una variación de los tiempos de respuesta del sistema, así como un aumento en el volumen de la fuente de datos luego de incorporada la propuesta.

Palabras Clave: calidad de datos, evaluación de calidad, Código de Autenticación de Mensajes (MAC).

ABSTRACT
The decision-making processes provide information of vital importance to any organization that intends to be competitive. Computer systems are fundamental for a good decision-making process. In this context, it is essential that the information that supports the decisions come from data that have the adequate quality, identifying integrity as one of the most important attributes. The principle of security is followed as an attribute of quality of the systems, particularizing in the security and quality of the data. The main contribution lies in the use of message authentication code for evaluating the quality that exhibits the data stored in a vulnerable data source. The results obtained in the experimentation reveal a variation of system response times, as well as an increase in the volume of the data source, after the proposed solution is incorporated.

Keywords: *data quality, quality assessment, messages authentication code.*

I. 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. Permiten organizar el procesamiento de información, agilizando así los procesos de gestión del negocio[1]. Los volúmenes de datos sobre los que se sustentan los sistemas van en aumento, de igual manera ocurre con la dependencia de las organizaciones con éstos en aras de elevar su competitividad y confiabilidad.

La información almacenada establece la historia y el presente de las organizaciones, constituyendo el asiento fundamental sobre el que se soporta la toma de decisiones de las mismas para su mejor desempeño [2,3]. Para realizar un proceso de toma de decisiones efectivo es imprescindible contar con sistemas que se ajusten a las características de los procesos y que sean confiables[2]. Un sistema confiable requiere que sus datos tengan la calidad requerida por los usuarios de forma tal que sea posible obtener información útil y confiable.

Por lo anterior, se requiere monitorear la calidad de los datos almacenados, con el objetivo de establecer un proceso que permita identificar y corregir anomalías en ellos, evitando de esta forma que los procesos de toma de decisiones estén soportados sobre información incorrecta. Dasu et al. (2003) plantean que es habitual que un gran porciento de los datos almacenados en una base de datos posea problemas de calidad [4]. La tarea de medir la calidad de los datos sin ayuda de herramientas y métodos especializados puede ser complicada en el entorno de las bases de datos. Estas almacenan grandes volúmenes de datos, por lo cual sería necesario mucho tiempo y esfuerzo para llevar a cabo el análisis de los mismos.

Como resultado de la necesidad de garantizar la calidad de la información surge una nueva área de investigación centrada en el estudio de la calidad de datos. La calidad de datos queda definida en la ISO/IEC 25012:2008, como el grado en que los datos satisfacen las necesidades de los usuarios[5]. Está compuesta por distintos aspectos conocidos como dimensiones de calidad, las cuales permiten realizar un estudio en forma más detallada de la calidad de datos. En la actualidad, existen varias técnicas a través de las cuales la calidad de datos puede ser evaluada entre las que se puede hacer mención del uso de metadatos y métricas [6, 7, 8, 9, 10].

Diversos son los factores por los cuales los datos de una aplicación pueden presentar mala calidad. Los investigadores que estudian la temática se centran en la entrada de los datos y la procedencia de los mismos en caso de que provengan de varias fuentes de información [11, 12]. Sin embargo, no son estos los únicos factores que propician que los datos puedan perder sus parámetros de calidad. Hasta el momento ninguno de los estudios toma en cuenta la posibilidad de que los datos puedan ser **ensuciados** desde un contexto de (in)seguridad (Del inglés: *(un) security*). Existe un claro escenario desde el punto de vista de la seguridad, en el cual la calidad de los datos puede ser afectada. Los datos almacenados en una base de datos o cualquier otro almacén de datos pueden estar expuestos a modificaciones impropias por personal no autorizado y con altos privilegios. Ante la ocurrencia de este fenómeno, puede ser comprometida la calidad de los datos almacenados y con ella la información utilizada en la toma de decisiones. Por todo lo anterior, es posible afirmar que existe una relación entre la calidad de datos y la seguridad.

Disímiles han sido los métodos desarrollados enfocados a medir y garantizar la calidad de los datos. Cada uno de estos permite identificar la ocurrencia de los problemas de calidad de datos en cada uno de los contextos estudiados anteriormente. Sin embargo, en ninguno de estos es contemplado el escenario en el que por violación de la integridad de los datos en el contexto de seguridad mencionado anteriormente estos pierdan los niveles de calidad requeridos. Esto implica que los métodos desarrollados hasta el momento sean vulnerables a este contexto. De igual manera que el dato es vulnerado, el medio mediante el cual es verificada su calidad pudiera estar siendo modificado por el atacante. Esto provoca el uso de datos erróneos sin ser detectado.

A partir de lo anterior, la principal contribución radica en el uso de Código de Autenticación de Mensajes (MAC) como mecanismo criptográfico resistente a ataques, para medir la calidad de los datos en las dimensiones exactitud, completitud y consistencia de la calidad de los datos. Los resultados obtenidos en la experimentación demuestran que el método propuesto es resistente al contexto de seguridad antes referido. Se evidencia una variación de los tiempos de respuesta del sistema, al igual que un aumento del volumen de los datos. Por otra parte, se demuestra la eficacia del método detectando modificaciones impropias en los datos.

II. MÉTODOS

Varios han sido las definiciones para la calidad de datos establecidas por ISO (2008), Oliveira, et al. (2005) y Guerra García, et al. (2015) [5, 12, 13]. Sin embargo, todos coinciden en que la calidad de datos está compuesta por distintas características, comúnmente conocidas como

dimensiones de calidad. Cada dimensión refleja un aspecto distinto en cuanto a la calidad de los datos.

Los artículos estudiados de Hongwei, et al (2014), Neely (2005), Yang W (1997), Wang et al (1995) identifican un núcleo de dimensiones comunes a la mayoría de los autores [14, 15, 16, 17]. La ISO/IEC 25012:2008[5] define un modelo de calidad de datos conformado por 15 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. Nótese la existencia de un conjunto de dimensiones resaltadas, las cuales tienen correspondencia con los dos puntos de vista, tal y como se muestra en la tabla 1. Las dimensiones objetivo en el presente artículo son: exactitud, completitud y consistencia, debido a la relación existente entre estas y el atributo de seguridad de la integridad.

Tabla 1. Dimensiones de la calidad de datos
Según los diferentes enfoques

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

La comunidad científica relacionada con el tema, identifica los problemas de calidad de datos como anomalías, errores o incluso suciedad[12]. Son múltiples las investigaciones realizadas en torno a los problemas de calidad de datos, en las cuales han sido definidos cada uno de los problemas identificados[18][19][12].

Oliveira plantea que los problemas de calidad de datos relacionados a los atributos y las filas pueden ser agrupados en cinco niveles fundamentales[12]:

- Único Atributo en una Única Fila, del inglés: Single Attribute of a Single Tuple(SAST)
- Único Atributo en Múltiples Filas, del inglés: Single Attribute in Multiple Tuples(SAMT, una columna)
- Múltiples Atributos en una Única Fila, del inglés: Multiple Attributes of a Single Tuple (MAST, una fila)
- una Única Relación, del inglés: Single Relation (SR), Múltiples Relaciones, del inglés: Multiple Relations (MR) y múltiples fuentes de datos, del inglés: Multiple Data Sources. (MDS)

Seguidamente, es ilustrado el dominio común de problemas de calidad de datos identificados en la bibliografía consultada. En la tabla 2 son mostrados los problemas siguiendo una clasificación basada en los niveles de granularidad antes mencionados[12].

Tabla 2. Problemas de calidad de datos

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

Cada uno de los problemas identificados es agrupado en los distintos niveles de granularidad en los cuales pueden surgir. 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.

La mayoría de los autores en la temática, analizan los problemas de calidad de datos fundamentalmente desde tres contextos diferentes. Un primer contexto cuando se corrigen anomalías existentes en una única fuente de datos. El segundo está dado cuando se realiza la migración de datos no estructurados a una fuente de datos estructurada. El tercero cuando se realizan procesos de integración de información proveniente de diversas fuentes en una única fuente datos[12]. Cada uno de los contextos anteriormente enunciados constituyen amenazas que desencadenan los problemas de calidad en los diferentes niveles mencionados anteriormente, los cuales producen afectaciones en las dimensiones de calidad de datos.

A pesar de que existe un amplio estudio de la calidad de datos en cada uno de estos tres contextos, lo que ha permitido el desarrollo de métodos de medición de calidad, existe un entorno no mencionado, hasta el momento, en la bibliografía consultada. Este contexto está relacionado con la amenaza real de que un atacante realice modificaciones en los datos y en las evidencias que permitirán medir el nivel de calidad del mismo accediendo a esta de forma impropia.

La información constituye el elemento fundamental en el proceso de toma de decisiones de las entidades, por ello es de vital importancia conocer los niveles de calidad con que cuenta. Para ello es imprescindible llevar a cabo un proceso de evaluación de la calidad sobre los datos de la organización. Un proceso de evaluación es la manera a través de la cual pueden ser identificados datos relacionados a un elemento específico que permiten establecer criterios para determinar en qué medida estos elementos cumplen con los fines y objetivos establecidos. Diversas son las técnicas existentes orientadas todas a los procesos de medición y mejoramiento de la calidad de

los datos. A continuación, se hace mención de las técnicas más populares en el entorno de la evaluación de la calidad de los datos.

Uso de ontologías

El grado de calidad con que cuentan los datos puede ser evaluado a través del uso de ontologías. Generalmente las ontologías son utilizadas para describir el conocimiento de un dominio determinado [19]. Estas permiten evaluar los problemas de calidad de datos mediante el desarrollo sistemático de métodos automatizados válidos y confiables [20]. El uso de ontologías en función de la evaluación de la calidad provee un conjunto de beneficios [21]. Está escrita en un lenguaje formal, es capaz de representar la semántica, provee un vocabulario compartido para discutir la calidad de los datos y es suficientemente riguroso para ser utilizado directamente en algoritmos y computadoras [21]. Sin embargo, esta técnica es susceptible a un contexto de seguridad. Hasta el momento no se evidencia el empleo mecanismos de protección de la ontología en este contexto.

Uso de metadatos

Una de las técnicas pioneras utilizadas en el control de la calidad de los datos es el uso de metadatos. Los metadatos son archivos que almacenan las características descriptivas de algún dato o recurso. Contienen: descripción detallada del dato, su composición, su origen, entre otras características relevantes [6]. En el caso del uso de esta técnica el dato es almacenado junto con su metadato. Ante la materialización de una modificación del dato mediante la violación de integridad del mismo, no existirá correspondencia entre este y el metadato asociado. Esta técnica no es resistente a un contexto de seguridad, de no estar protegido el metadato, este puede ser modificado por el atacante de igual manera que el dato. De ocurrir esto se afectaría la calidad del dato sin ser detectada la ocurrencia de este fenómeno.

Uso de métricas

Una de las técnicas más utilizada para evaluar la calidad con que cuentan los datos es el uso de métricas. Numerosos autores proponen el uso de métrica para llevar a cabo procesos de evaluación de la calidad [7, 8, 9, 21]. Las métricas de evaluación son funciones que pueden ser ajustadas a situaciones específicas de evaluación [23, 24]. Cada una de las dimensiones de la calidad de datos puede ser medida mediante el uso de métricas, estas definen la forma en la que son medidos los factores de calidad [7, 22]. Las funciones por las cuales pueden estar definidas las métricas pueden realizar comparación entre datos, cálculo de la distancia entre dos valores, etc [18]. Las aplicaciones de estas funciones proveen un resultado numérico que permiten determinar el rango de ocurrencia de determinados problemas en la dimensión medida [10].

Los estudios referentes a la calidad de los datos, así como a los problemas de calidad de datos se han centrado en tres contextos fundamentales anteriormente enunciados. Los métodos desarrollados hasta el momento para evaluar la calidad de los datos se enmarcan en dichos contextos. En correspondencia con la situación actual de los ataques a los sistemas informáticos, no se evidencia un estudio realizado desde un contexto de seguridad, en el cual producto a una amenaza de seguridad la calidad de los datos puede quedar total o parcialmente comprometida [25, 26].

El contexto de seguridad queda definido por los datos esenciales con que cuenta una organización, siendo de interés garantizar que estos sean confiables. Por lo anterior, existe la posibilidad de que los administradores de los sistemas o cualquier otro personal no autorizado realicen modificaciones sobre los datos. En este caso se incurriría en la violación de uno de los principales principios de la seguridad de la información, se violaría la integridad de los datos comprometiendo así la calidad de los mismos. La situación anterior evidencia la existencia de una relación entre la integridad de la información desde el entorno de la seguridad y la calidad de la misma. Lo cual permite el empleo de mecanismos de seguridad para llevar a cabo procesos de evaluación de la calidad. A su vez, el contexto propuesto hace vulnerable cada una de las técnicas de evaluación de calidad abordadas anteriormente, ya que en ninguna de estas se contempla el escenario en que la información es modificada mediante la violación de la integridad desde el entorno de seguridad. Ante la materialización de la amenaza ninguna de las técnicas desarrolladas hasta el momento detectaría la ocurrencia de la misma, arrojando así un resultado erróneo acerca del nivel de calidad con que cuentan los datos evaluados.

Según Mario Piattini, la auditoría informática es: "el proceso de recoger, agrupar y evaluar evidencias para determinar si un sistema informatizado salvaguarda los activos, mantiene la integridad de los datos, lleva a cabo eficazmente los fines de la organización y utiliza eficientemente los recursos" [27]. Por las variantes tradicionales de auditoría es muy difícil detectar la pérdida de la calidad de los datos producto a estos ataques activos en los cuales es

violada la integridad del mismo, ya que estos no generan una traza mediante la cual los procesos de auditoría puedan detectar el error.

Por lo anterior, el propósito de la presente investigación es proporcionar un método resistente al contexto de seguridad que permita medir la calidad con que cuentan los datos, ya que no se ha encontrado ninguna evidencia acerca de la existencia de un método resistente a este contexto. Para ello se realizó un estudio acerca de los mecanismos para garantizar integridad de los datos existentes independientemente de los incorporados en los sistemas gestores de bases de datos. En la figura 1 se ilustran los mecanismos estudiados clasificados según el estado final del mensaje.

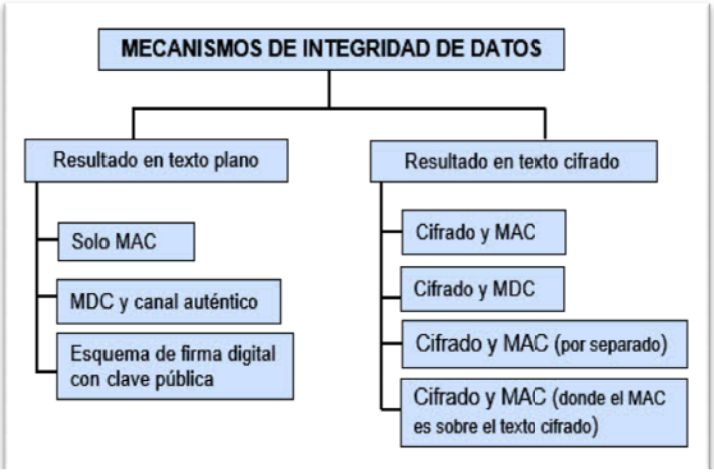

Fig.1. Mecanismos de control de integridad según estado final del mensaje

En el presente artículo, se decide emplear un mecanismo de Solo MAC, ya que frecuentemente son empleados en aplicaciones en las cuales garantizar la integridad de datos constituye un objetivo. Generan un valor MAC asociado al dato mediante el uso de funciones de *hash* criptográficamente seguras. Seguidamente son descritas las características del mecanismo empleado.

Los algoritmos MAC requieren que tanto el emisor del mensaje como el receptor compartan un clave secreta. Para una mayor comprensión en la figura 2 se ilustra el procedimiento del algoritmo MAC[28].

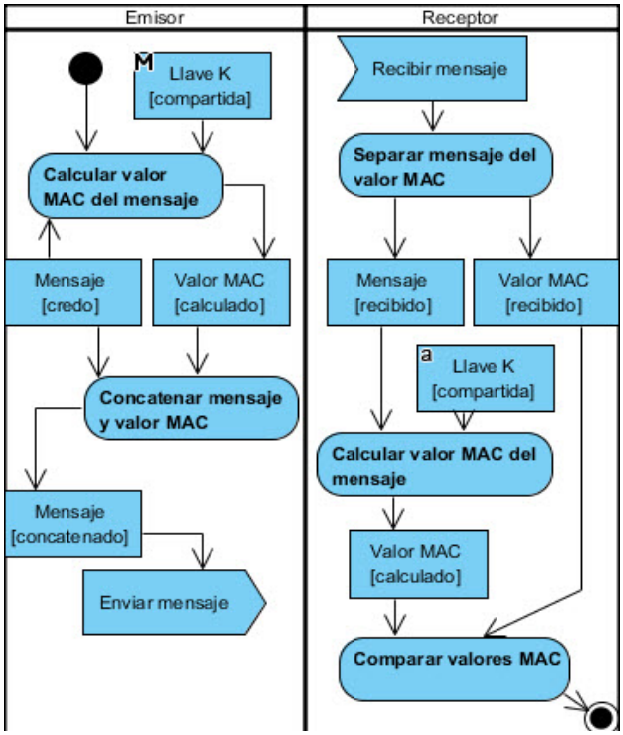

Fig.2. Integridad de dato mediante MAC

El emisor del mensaje calcula un MAC sobre este empleando una clave secreta. La clave secreta debe ser compartida entre el emisor y el receptor del mensaje. Seguidamente el emisor envía el mensaje y el valor MAC, el mensaje y su valor MAC se encuentran divididos por algún identificador previamente acordado por las partes. Una vez que al receptor lo recibe, 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. En caso de que ambos valores sean iguales, el receptor tiene total seguridad de que el mensaje es auténtico e íntegro. Por tal motivo, se puede garantizar que el mensaje ha sido generado por una parte que conoce la clave compartida y no fue alterado[28][29].

Esta técnica es aplicable al contexto de las bases de datos. Una vez que es almacenado el mensaje, el mismo puede ser consultado siempre que se requiera. Para el caso en el que se quiera verificar la integridad del mismo se consulta su valor de integridad en la base de datos y conociendo la clave secreta es posible verificar su integridad.

Propuesta

El objetivo de esta sección es describir los elementos fundamentales del método empleado para 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. El método de medición queda definido por la siguiente expresión (1):

$$MC \equiv (MD, F, T: (MD, F, ED) \rightarrow ED') \quad (1)$$

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 de la siguiente manera como se aprecia en la expresión (2):

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

Donde se identifica como d el dato vulnerable al cual se medirá la calidad. La función es la encargada de generar el valor de evidencia de la forma a partir del dato d, mientras que la función se define como en la expresión (3). Para el caso en particular de este artículo la función estará compuesta por el algoritmo MAC descrito en la sección anterior.

$$f_v(x, d) = \begin{cases} 0, & x \neq f_G(d) \\ 1, & x = f_G(d) \end{cases} \quad (3)$$

Por último, $T: (MD, F, ED) \rightarrow ED'$ 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. Como resultado de estas transformaciones el esquema de datos puede tener una nueva columna en la tabla donde se encuentra el dato vulnerable, una nueva fila o una nueva tabla entre otros.

III. RESULTADOS

Para la validación del método propuesto fue llevado a cabo un proceso de experimentación con dos objetivos fundamentales. Un primer objetivo consiste en evaluar la eficacia del método detectando violaciones de integridad en los datos. El segundo objetivo propuesto está dado por la realización de un análisis del comportamiento de los tiempos de respuesta del sistema luego de incorporada la propuesta.

Para llevar a cabo la experimentación fue utilizada la biblioteca *bouncy castle*, esta proporciona el conjunto de métodos criptográficos necesarios [30]. Fue empleado un mecanismo de Solo MAC para llevar a cabo el cifrado de los datos vulnerables. Dicho mecanismo recibe en dato junto con un atributo llave a partir del cual la evidencia es generada. El cálculo del valor MAC se realiza empleando la función HMac-SHA3-512. Finalmente, la evidencia generada es almacenada en la base de datos.

La experimentación fue realizada en el Módulo de Archivo Histórico del SIGENU (Sistema de Gestión de la Nueva Universidad). El sistema SIGENU es el encargado de gestionar la información referente a los procesos docentes en los centros de educación superior en Cuba. El módulo de Archivo Histórico se centra en la información docente de los egresados, tanto los graduados como aquellos estudiantes que causaron baja definitiva del centro de estudios. La principal función de este módulo es hacer persistir la información en el tiempo de manera tal que pueda ser consultada en tiempos futuros.

Para ello fue identificado como dato sensible la nota de los estudiantes en cada una de las asignaturas cursadas durante su trayecto. Para cada una de las notas a proteger es generada 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).

El Módulo de Archivo Histórico consta de 4 funciones fundamentales que se encargan del procesamiento de esta información.

F1-*getMatriculatedSubjectByStudent*, función a través de la cual son listadas las asignaturas cursadas por el egresado.

F2- *addSubjectByStudent*, función mediante la cual se matricula una asignatura cursada a un egresado.

F3- *updateMatriculatedSubject*, esta función permite la actualización de los datos en las asignaturas matriculadas de un egresado.

F4- *addSubjectByStudentLote*, a través de dicha función se puede realizar la matrícula de más de una asignatura a más de un egresado.

Una vez integrada la propuesta al sistema se procedió a registrar las notas de las asignaturas matriculadas de un estudiante generándose para cada una la evidencia correspondiente. Seguidamente se modificaron las notas de parte de las asignaturas matriculadas accediendo directamente a la base de datos, con el objetivo de evaluar la eficacia del método ante ataques. Al consultar 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.

Impacto en el rendimiento

Para la valoración del impacto de la solución en el rendimiento del sistema se realizó un proceso de experimentación con el objetivo de analizar 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 el método con el uso del mecanismo de solo MAC. Los experimentos fueron realizados sobre la base de datos del Sistema de Archivo Histórico del SIGENU.

Para comprobar si la utilización de dicho mecanismo de integridad produce cambios en los tiempos de ejecución del sistema fue realizada una prueba de hipótesis. A continuación, se describe la prueba realizada.

Las hipótesis evaluadas fueron:

H_0: T_s-T_{mac} =0

H_1: T_s-T_{mac} ≠0

Con α= 0.05

Donde:

H_0: Hipótesis nula

H_1: Hipótesis alternativa

α: Máximo nivel de riesgo aceptable para rechazar una hipótesis nula verdadera

T_s: Tiempo de respuesta del sistema sin el método incorporado

T_{mac}: Tiempo de respuesta del sistema luego de incorporar el método utilizando solo MAC

Fue utilizada la prueba no paramétrica de Mann-Whitney. La selección se realizó sobre la base de no existir evidencia acerca de que 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 H0 es verdadera. Los resultados obtenidos se muestran en la tabla 3.

Tabla 3. Valores obtenidos de la experimentación

Función	Valores p	Promedio de los tiempos		Mediana de los tiempos	
		Sin método	Con método	Sin método	Con método
F1	0.370	70.8	72.2	70	71
F2	0.272	57.9	58.5	58	59
F3	0.064	2.4	2.9	2	3
F4	0.155	19.5	21.4	18.5	19.5

Impacto en el tamaño de la base de datos

Otro de los aspectos fundamentales que valoran las organizaciones ante la implantación de nuevas soluciones es el incremento que implican las mismas en las bases de datos sobre las que se sustentan. Por tal motivo a continuación, se realiza un análisis del crecimiento de la base de datos del sistema una vez que es incorporada la propuesta de solución. El incremento del tamaño está dado por la expresión 4.

$TA = CE * CA * TR$ (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 agregado, el valor se encuentra asociado al método que se emplee

Anualmente el promedio de graduados en el centro 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 contara con un valor de redundancia asociado a la evaluación de la misma. La tabla 4 muestra los resultados obtenidos.

Tabla 4. Incremento de la fuente de datos

Mecanismo de integridad	CE	CA	TR	TA
Solo MAC	1400	69	520 bits	5 MB

IV. DISCUSIÓN

Los principales aportes de esta investigación consisten en la identificación de un nuevo contexto de seguridad en el cual pueden surgir los problemas de calidad de datos. Se define un método de evaluación de calidad de datos basado en principios de seguridad, el cual mediante el empleo de código de autenticación de mensajes permite identificar la pérdida de calidad producto a violaciones de integridad de los datos.

El nuevo método de evaluación permite identificar la pérdida de calidad de datos como resultado de la violación de la integridad del dato. Por ello, que en aras de obtener un mejor resultado se proponen las siguientes recomendaciones:

La función a emplear para generar el cifrado debe ser segura, no se recomienda utilizar una función que haya sido vulnerada como es el caso de las funciones del tipo MD (tales como MD5).

Se recomienda el uso de funciones de hash criptográficamente seguras, libres de colisiones como es el caso de la función HMac-SHA3-512 empleada en el proceso de experimentación.

En la actualidad existen ataques reconocidos sobre textos cifrados en los cuales los atacantes corrompen la llave empleada en el cifrado. Es por ello que la clave empleada debe ser bien conformada y protegida de manera adecuada.

A partir de los resultados obtenidos en la experimentación, queda demostrada la eficacia del método propuesto detectando modificaciones impropias sobre los datos. Una vez finalizados los experimentos se puede afirmar que el método propuesto es resistente al contexto de seguridad expuesto en secciones anteriores. Teniendo en cuenta que la propuesta queda definida por parámetros independientes del entorno de experimentación, es posible afirmar que la misma es generalizable y aplicable a cualquier entorno en el cual se realice la manipulación de datos. El método propuesto podrá ser incluido en cualquier sistema de gestión de datos en el cual se identifique información sensible que demande determinados niveles de calidad.

Por otra parte, los resultados arrojados del análisis estadístico para el empleo de la técnica de Solo MAC evidencian que el uso de esta técnica no implica un aumento significativo en los tiempos de respuesta del sistema. Para el caso del incremento de la base de datos se obtiene que experimentará un crecimiento de 5 MB anuales lo que representa solamente el 0.9% del tamaño original.

V. CONCLUSIONES

1. La correcta toma de decisiones es la clave fundamental para el crecimiento de las organizaciones. Estas se sustentan sobre la base de contar con información con la calidad requerida por los directivos. Siendo los datos los protagonistas del éxito de las organizaciones es vital garantizar que cuenten con el nivel de calidad adecuado.
2. En el presente artículo se realiza un estudio de la calidad de datos desde el punto de vista de la seguridad, en el cual la materialización de modificaciones impropias afectaría la calidad de los datos almacenados. A partir del estudio realizado, se define un nuevo contexto de seguridad en el cual pueden surgir los problemas de calidad de datos. El aporte principal de este trabajo está dado por la definición de un método que, mediante el empleo de códigos de autenticación de mensajes permite evaluar la calidad de los datos en el nuevo contexto definido.
3. Se realizó un análisis del rendimiento del sistema luego de incorporado el método propuesto, el cual demostró que no implica aumentos significativos en los tiempos del sistema. De igual modo, se considera que el impacto en el tiempo de respuesta no aparenta ser alto en comparación con el resultado de calidad que se alcanza. Finalmente, el análisis del comportamiento del volumen de datos arrojó como resultado que se experimentará un crecimiento anual de un 0.9% del tamaño original.

VI. REFERENCIAS

1. Ramírez JL, Vega O. Sistemas de información gerencial e innovación para el desarrollo de las organizacione. Télématique. 2015 (14). ISSN 1856-4194.
2. C. Samitsch D. Data Quality and its Impacts on Decision-Making. En: How Managers can benefit from Good Data. Germany: Gabler Verlag. ISBN 9783658081997.
3. Subiela Durá S. Sistemas de Información BI: Estado actual y herramientas de software libre.2011. [Citado: 25 de octubre del 2018]. Disponible en: openaccess.uoc.edu/webapps/o2/bitstream/10609/8175/1/Sduras_TFM_0611.pdf..
4. Dasu T, al e. Data quality through knowledge engineering. Ninth International Conference on Knowledge Discovery and Data Mining. 2003. [Citado: 26 de octubre del 2017]. Disponible en: https://www.myhuiban.com/conference/136.
5. ISO. Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. Suiza: ISO/IEC, ISO/IEC FDIS 2008.
6. Verbo E, Caballero I, Pérez R, et al. Una Metodología Basada en ISO/IEC 15939 para la Elaboración de Planes de Medición de Calidad de Datos.2008 En: 13th Conference on Software Engineering and Databases. JISBD. p. 253–64. ISBN 978-84-9749-486-1.
7. Zaveri A, Rula A, Maurino A, et al. Quality Assessment for Linked Open Data: A Survey. Semantic Web journal. 2016;7(1):63-93. ISSN 1570-0844.
8. Shankaranarayanan G, Ziad M, Wang RY. Preliminary Study on Data Quality Assessment for Socialized Media. China Science and Technology Resources. 2012;44(2):72-9. ISSN 1673-4874.
9. Aljumaili M. Data Quality Assessment: Applied in Maintenance. Luleå, Swedish: Luleå University of Technology; 2016. ISBN 978-91-7583-521-1.
10. Sadiq S. HandBook of Data Quality: Research and Practice. New York: Springer Heidelberg; 2013. ISBN 3642362567. DOI
11. Sidi F, Panahy PHS, Affendey LS, et al. Data Quality:A Survey of Data Quality Dimensions: IEEE; 2013. ISBN 978-3-540-43349-1.
12. Oliveira P, Rodrígues F, Henriques P. A Formal Definition of Data Quality Problems.2005 En: International Conference on Information Quality. Cambridge, MA, USA. MIT. p. ISBN 1-59593-160-0.:
13. Guerra García C, Hernández Domínguez V, Caballero I, et al. Selecting web functionalities versus data quality dimensions: A first approach.2015 En: 5th International Symposium on Data-driven Process Discovery and Analysis Graz, Austria. Springer. p. ISBN 978-3-319-74160-4. Disponible en:
14. Hongwei Z, Stuart EM, Lee YW, et al. Data and information quality research: Its evolution and future. Londres: Taylor & Francis Group, LLC; 2014. ISBN 978-3-319-13289-1.
15. Neely MP. The Product approach to data quality and ftness for use: A Framework for analysis,.2005 En: 10th International Conference on Information Quality. Cambridge, MA, USA. MIT. p. ISBN ISBN 1-59593-160-0.
16. Lee YW, Strong DM, Wang RY. Data quality in context. Communications of the ACM. 1997;40(5):103--10. ISSN 0001-0782.

17. Wang RY, Reddy MP, Kon HB. Toward quality data: an attribute-based approach. Decision Support Systems. 1995 (13):349–72. ISSN 0167-9236.
18. Chen H, Hailey D, Wang N, et al. A review of data quality assessment methods for public health information systems. International Journal of Environmental Research and Public Health. 2014;11(5). ISSN 1661-7827.
19. Barchini GE, Alvarez MM, Palliotto D, et al. Evaluación de la calidad de los sistemas de información basados en ontologías.2009 En: IX Congreso ISKO-España. Valencia. Editorial UPV. p. ISBN 9788483633984
20. Liaw S-T, de Lusignan S, Taggart J, et al. Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. International Journal of Medical Informatics. 2012;81(1):10-24. ISSN 1386-5056.
21. Pipino LL, Lee YW, Wang RY. Data Quality Assessment. Communications of the ACM. 2002;45(April). ISSN 1460-2466.
22. Laranjeiro N, Soydemir SN, Bernardino J. A Survey on Data Quality: Classifying Poor Data.2015 En: 21st Pacific Rim International Symposium on Dependable Computing. Zhangjiajie, China. IEEE. p. ISBN 978-1-5090-5652-1.
23. Valverde C, Marotta A, Vallespir D. Análisis de la Calidad de Datos en Experimentos en Ingeniería de Software.2012 En: XVIII Congreso Argentino de Ciencias de la Computación. Bahía Blanca, Argentina. Universidad Nacional del Sur. p. ISBN 978-987-1648-34-4
24. Sains F, Teknologi UMT. Data Investigation: Issues of Data Quality and Implementing Base Analysis Technique to Evaluate Quality of Data in Heterogeneous Databases. Journal of Theoretical and Applied Information Technology. 2012;45(1). ISSN 1992-8645.
25. Centeno Ureña FJ. Ciberataques, la mayor amenaza actual.2015. [Citado: 25 de octubre del 2018]. Disponible en: http://www.ieee.es/temas/ciberseguridad/2015/DIEEEO09-2015.html
26. Medero Sánchez G. La ciberguerra: los casos de Stuxnet y Anonymous. Derecom.11(1). ISSN 1988-2629.
27. Velthuis MP. Auditoría de Tecnologías y Sistemas de Información: Ra-Ma; 2008. ISBN 978-84-7897-849-6.
28. Preneel B. The State of Cryptographic Hash Functions. En: Lectures on Data Security, Modern Cryptology in Theory and Practice, Summer School, Aarhus, Denmark, July 1998 London, UK: Springer-Verlag; 1999. p. 158–82. ISBN 3-540-65757-6
29. Menezes AJ, van Oorschot PC, Vanstone SA. Handbook of Applied Cryptography. EEUU CRC Press 1996. ISBN 9780849385230.
30. Knudsen J. Wireless Java: Developing with Java 2. Michigan: Universidad de Michigan: Apress; 2007. ISBN 0672320959.