UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

Título: Estrategia de seguridad para soluciones de Almacenes de datos

Trabajo de Diploma para optar por el título de Máster en Informática Avanzada

Autora:

Ing. Yuneimy Tellez Pérez

Tutores:

Msc. Ruth Yurina Vega Cutiño

Dr. Juan Pedro Febles Rodríguez

La Habana, diciembre de 2016

Dedicado especialmente *a mi pequeña Alejandra por ser mi inspiración, mi fuerza para lograr cada meta en mi vida,*

A Ale mi maravilloso esposo, por ser tan especial y apoyarme en todo cuanto fue preciso para que hoy alcanzara este sueño,

A mis madres Maribel y Midelma, quienes juntas con mucho amor y cariño me han convertido en la persona que soy hoy,

A mi abuela Isabel por ser más que una madre, por esa fuerza que tiene y me inspira,

A mi abuelo Pedro, que me dio siempre un amor tierno e infinito y quien me enseñó el valor que tiene la dedicación y el sacrificio cuando perseguimos lo que queremos,

A mi padre Bernardo, por ser mi ejemplo, mi guía y consejero,

A mi hermano, con ese corazón tan noble que siempre está ahí cuando lo necesito,

A mis tíos Norge y Ernesto, que siempre han sido como mis padres,

A toda mi familia.

Agradezco especialmente a *mis tutores Ruth y Febles por dedicarme todo el tiempo que pudieron aun cuando no lo tienen, por trasmitirme sus conocimientos y por ser personas tan sencillas y nobles,*

A mi pequeña hija y a mi esposo,

A mis madres,

A mi padre,

A mis abuelos,

A mi amiga y hermana Doris, que me ha demostrado que la amistad verdadera sí existe,

A mis amigos Yonelbys, Keimer, Garnache, Paula y Yayi que tanto me apoyaron,

A mis compañeros de trabajo del antiguo departamento "Almacenes de datos" de la Facultad 6,

A mis amigos y compañeros de trabajo en Midas,

A todos los que estuvieron al tanto de cómo iba este trabajo, mis más sinceros agradecimientos.

DECLARACIÓN DE AUTORÍA

Declaro que soy la única autora de este trabajo y autorizo a la Dirección General de Producción de la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

_____	_____
Firma del maestrante	Firma de supervisores

RESUMEN

Los almacenes de datos en la actualidad, representan una solución importante para el análisis del gran volumen de datos que se genera en cualquier empresa u organización, y constituyen un soporte para la ayuda a la toma de decisiones, la administración y el control de cualquier institución. Debido a esto, se le acredita un gran valor a la confidencialidad de la información que integran los almacenes de datos. La seguridad en todo el ciclo de desarrollo de este tipo de soluciones es un requisito fundamental. En el presente trabajo se expone una estrategia de seguridad para soluciones de almacenes de datos del centro de Tecnologías y Gestión de Datos de la Universidad de las Ciencias Informáticas.

La estrategia de seguridad propuesta representa una guía de acciones para los desarrolladores de almacenes de datos, con el propósito de lograr sistemas seguros interviniendo en cada uno de los subsistemas que se desarrollan. Las acciones organizadas en una espiral de mejora continua contribuyen a elevar la calidad de estas soluciones. Se abordan principios de seguridad, mecanismos de seguridad y una propuesta de arquitectura que permitirá la implementación de los elementos de seguridad que se exponen. Además, se corroboró la eficacia de la aplicación de la estrategia de seguridad en el almacén de datos: “Sistema de análisis estadísticos para los procesos eleccionarios en Cuba”; donde se aplicaron buenas prácticas de seguridad para cada subsistema presente. Finalmente se evaluó la estrategia de seguridad, obteniéndose resultados que califican la propuesta con una alta probabilidad de éxito.

Palabras clave: almacén de datos, arquitectura, estrategia de seguridad, mecanismos de seguridad, seguridad.

ÍNDICE

INTRODUCCIÓN 1

CAPÍTULO I. MARCO TEÓRICO REFERENCIAL DEL PROCESO DE IMPLEMENTACIÓN DE MECANISMOS DE SEGURIDAD EN EL DESARROLLO DE ALMACENES DE DATOS 6

1.1 Almacenes de datos. Arquitectura 6

1.2 Almacenes de Datos. Factores de riesgos y vulnerabilidades 8

1.3 Almacenes de Datos. Requisitos de Seguridad 11

1.4 Propuestas de seguridad existentes para almacenes de datos 13

1.5 Principios y mecanismos de seguridad 15

1.6 Estado de la seguridad en soluciones de AD del centro DATEC 19

1.7 Estándar ISO/IEC 27001 20

1.8 Acciones a seguir para la seguridad de los almacenes de datos 20

1.9 Plataforma de Inteligencia de Negocio. Suite de Pentaho 22

1.9.1 Pentaho BI Server 23

1.10 Mecanismos de autenticación 24

1.10.1 Servicio Central de Autenticación 25

1.10.2 Protocolo Ligero de Acceso a Directorios 26

1.11 Arquitectura Orientada a Servicios 27

1.12 Gestión del conocimiento 28

1.13 Características de la estrategia científica 30

1.14 Conclusiones parciales 31

CAPÍTULO II. DESCRIPCIÓN DE LA ESTRATEGIA DE SEGURIDAD PROPUESTA PARA SOLUCIONES DE ALMACENES DE DATOS 33

2.1 Propuesta de solución 33

2.2 Estrategia de seguridad para almacenes de datos 34

2.2.1 Acciones de la estrategia de seguridad 35

2.2.2 Principios de la estrategia de seguridad 39

2.2.3 Propuesta de arquitectura de seguridad 39

2.2.4 Mecanismos de seguridad a nivel de subsistema 41

2.2.5 Conclusiones parciales 47

CAPÍTULO III. APLICACIÓN Y EVALUACIÓN DE LA ESTRATEGIA PROPUESTA 48

3.1 Aplicación de la estrategia de seguridad 48

3.2 Métodos seleccionados para validar la propuesta 54

3.3 Etapas presentes en la validación 55

3.3.1 Elección de expertos 55

3.3.2 Elaboración y lanzamiento de cuestionarios....57
3.3.3 Realización de pruebas de concordancia....58
3.3.4 Resultados de la evaluación....59
3.4 Conclusiones parciales....61
CONCLUSIONES....62
RECOMENDACIONES....62
REFERENCIAS BIBLIOGRÁFICAS....63
GLOSARIO....67
ANEXO 1. Formatos de ejemplos....68
ANEXO 2. Diagramas de componentes....71
ANEXO 3. Encuesta. Nivel de experiencia de los expertos....73
ANEXO 4. Encuesta. Medidores de los resultados de la investigación.... - 74 -
ANEXO 5. Encuesta. Valoración de los resultados de la investigación....76
ANEXO 6. Resultados de la evaluación. Pesos otorgados....77

INTRODUCCIÓN

En esta nueva era tecnológica, la acumulación de la información es común en la mayoría de las empresas, organizaciones e instituciones; en las cuales no es suficiente el uso de los sistemas operacionales[1] para satisfacer sus necesidades de información (Berzal Galiano, 2016). Con el objetivo de tomar decisiones factibles partiendo de los datos históricos disponibles en fuentes heterogéneas, se hace necesario realizar análisis exhaustivos, consultas y obtención de la información en el menor tiempo posible.

Una de las áreas de investigación que soluciona el problema planteando es la Inteligencia de Negocio (BI por sus siglas en inglés: *Business Intelligence*) mediante los Sistemas de Apoyo a la Toma de Decisiones (DSS por sus siglas en inglés: *Decision Support Systems*). Una definición ampliada es la que se propone en The datawarehouse Institute[2]: *"…BI abarca los procesos, las herramientas, y las tecnologías para convertir datos en información, información en conocimiento y planes para conducir de forma eficaz las actividades de los negocios. Abarca las tecnologías de datawarehousing, los procesos en el back end, consultas, informes, análisis y las herramientas para mostrar información y los procesos en el front end…".* Entre sus componentes fundamentales se encuentran los Almacenes de Datos[3] (AD).

El tema de almacenes de datos fue introducido a principio de la década de los años 90 por Bill Inmon, quien los definió como una colección de datos orientado a las materias, integrado, no volátil, que varía con el tiempo y que sirve de soporte al proceso de toma de decisiones (Trujillo, y otros, 2013). Se encuentran dentro de la línea evolutiva de las bases de datos, con el objetivo de lograr una mayor funcionalidad e inteligencia partiendo de la información generada en cualquier entorno empresarial. Su objetivo, es precisamente integrar datos corporativos en un único repositorio sobre el cual los usuarios pueden realizar consultas, informes y hacer análisis de los datos históricos que se generan en las empresas, organizaciones e instituciones.

Debido a que los AD contienen datos consolidados de múltiples fuentes, el acceso a los mismos por personas no autorizadas puede ser uno de los blancos fundamentales de una empresa (Security and the Data Warehouse, 2005). Como consecuencia, es imprescindible la incorporación de elementos de seguridad en el desarrollo de cualquier AD, para hacerlos confiables y accesibles solo por las personas autorizadas. Inconvenientemente, el desarrollo de estas soluciones ha estado encaminado fundamentalmente a la seguridad de elementos específicos y no se han considerado aspectos de seguridad que deben incorporarse en todas las etapas del ciclo de desarrollo de un almacén de datos (A UML 2.0/OCL Extension for Designing Secure Data Wharehouses, 2006). En la literatura se encuentran varias propuestas de seguridad para AD que tratan: el filtrado y cifrado de los datos antes de su almacenamiento en los AD (View Security as the Basic for DW Security, 2000), la propuesta de un modelo de control de acceso y auditoría para el modelado multidimensional de almacenes de datos (Access control and audit model for the

[1] OLTP: On-Line Transaction Processing.

[2] Organización fundada en 1995, promueve la construcción de almacenes de datos.

[3] También conocido como Data Warehouse.

multidimensional modeling of data warehouses, 2006), la construcción de un esquema en estrella seguro para soluciones de almacenes de datos (Building a secure star schema in data warehouse by an extension of rela-tional package from CWM, 2008), la aplicación de un enfoque basado en la arquitectura dirigida por modelos (MDA) para desarrollar almacenes de datos seguros (Applying an MDA-based approach to consider security rules in the development of secure DWs, 2009), el diseño de un enfoque híbrido para asegurar los almacenes de datos que integra modelos clásicos de seguridad y la aplicación del filtrado de los datos (An Improved Security Framework for Data Warehouse: A Hybrid Approach, 2010), entre otras propuestas.

El centro de Tecnologías de Gestión de Datos (DATEC) de la Universidad de Ciencias Informáticas (UCI), tiene entre sus propósitos desarrollar este tipo de soluciones para automatizar el control estadístico de cualquier institución del país; utilizando como metodología de desarrollo una adaptación que sigue el ciclo de vida propuesto por Kimball (González Hernández, 2013). Los proyectos que se desarrollan son de gran importancia e impacto social para las organizaciones y empresas del país, siendo la seguridad un requisito esencial, sin embargo, durante su desarrollo se presentan las siguientes problemáticas relacionadas con los elementos de seguridad:

- La metodología que guía el desarrollo de las aplicaciones en el departamento plantea que se deben tener en cuenta los requisitos de seguridad. Establece las actividades y fases que deben desarrollarse para implementar los requisitos de seguridad y los artefactos como resultado de su implementación. Inconvenientemente, no aborda los procedimientos necesarios que deben guiar dicha especificación e implementación.

- En la actualidad los AD desarrollados se centran en un grupo limitado de requisitos de seguridad. Generalmente se limitan a garantizar la seguridad en el diseño físico del despliegue de la solución, sin tener en cuenta la seguridad necesaria en todos los subsistemas que conforman la arquitectura utilizada: Fuente de datos, Procesos de Extracción, Transformación y Carga (ETL por sus siglas en inglés: Extract, Transform, Load), Almacén de datos y Visualización de los datos.

- La arquitectura definida no organiza la información almacenada en el AD, en relación con el valor de los datos y la gravedad de las amenazas a las que puedan estar sometidos. Esto imposibilita que se puedan definir diferentes mecanismos de seguridad teniendo en cuenta dicha relación.

- La herramienta Pentaho BI Server 6.1 utilizada para la visualización y el análisis de la información, brinda funcionalidades limitadas respecto a la seguridad de los datos; debido a que en su configuración básica: los mecanismos de autenticación se especifican de forma manual editando directamente los ficheros de configuración, las contraseñas de los usuarios no se validan eficazmente y tampoco existe la posibilidad de restringir el acceso mediante direcciones IP.

Todas estas limitaciones en cuanto a la seguridad de las aplicaciones realizadas pueden originar que se realicen intentos exitosos que comprometan la información confidencial, es decir, la información protegida por las organizaciones y que es importante para sus negocios y

operaciones. La divulgación, modificación y destrucción de la información pueden resultar en pérdidas económicas u otras consecuencias negativas para las instituciones.

Por lo expuesto anteriormente el **problema de la investigación** queda definido de la siguiente manera:

El proceso para implementar los mecanismos de seguridad no gestiona satisfactoriamente la seguridad de las soluciones de almacenes de datos del centro DATEC de la UCI.

La presente investigación tiene como **objeto de estudio**: proceso de implementación de mecanismos de seguridad en los almacenes de datos, enmarcado en el **campo de acción**: las estrategias de seguridad en las soluciones de almacenes de datos del centro DATEC de la UCI.

Para darle solución al problema de la investigación se define como **objetivo general**: desarrollar una estrategia para elevar la seguridad en las soluciones de almacenes de datos del centro DATEC de la UCI.

A partir del objetivo general se desglosan los siguientes **objetivos específicos**:

1. Establecer el marco teórico de la investigación relacionado con el proceso de implementación de mecanismos de seguridad en diferentes arquitecturas para la construcción de almacenes de datos.
2. Diagnosticar la situación actual de las estrategias de seguridad en las soluciones de almacenes de datos del centro DATEC de la UCI.
3. Desarrollar una estrategia de seguridad que permita implementar mecanismos de seguridad en todas las etapas del ciclo de desarrollo de un almacén de datos.
4. Evaluar la estrategia de seguridad desarrollada a través de la aplicación de técnicas multicriterios con el consenso de expertos.

Una vez definido e identificado el problema, sobre la base de la experiencia, se plantea la siguiente **hipótesis**:

Si se aplica una estrategia de seguridad en el proceso de implementación de los mecanismos de seguridad, entonces se elevará la seguridad de las soluciones de almacenes de datos del centro DATEC de la UCI.

Principales aportes de la investigación:

Aportes teóricos:

Estrategia de seguridad que incluye: las acciones a seguir en el proceso de implementación de mecanismos de seguridad para las soluciones de AD del centro DATEC de la UCI, principios de seguridad, propuestas de mecanismos de seguridad para cada subsistema presente y una propuesta de arquitectura de seguridad.

Aportes prácticos:

Propuesta para realizar modificaciones a la herramienta Pentaho BI Server 6.1, que solucione las limitaciones de seguridad presentes en su configuración básica.

Métodos teóricos: para el desarrollo de la investigación se utilizaron los métodos teóricos Histórico lógico, Hipotético deductivo, Modelación, Inductivo deductivo y el Analítico sintético.

Histórico lógico

El proceso de implementación de mecanismos de seguridad y las estrategias de seguridad orientadas a las soluciones de AD fue revisado desde un enfoque histórico lógico, al realizar el estudio del estado del arte lo cual permitió analizar las características, ventajas y limitaciones de los elementos estudiados.

Hipotético deductivo

La presente investigación siguió el método hipotético deductivo, porque partiendo de conocimientos generales del proceso de implementación de mecanismos de seguridad y de las estrategias de seguridad revisados en la literatura, se infirieron los elementos necesarios que debía incluir la estrategia de seguridad propuesta; según las características específicas del entorno en el cual está enmarcada la solución.

Modelación

La modelación se utilizó al representar gráficamente algunos de los elementos definidos en la estrategia de seguridad. El uso de este método científico permitió seguir una de las buenas prácticas de desarrollo de software al permitir la comprensión del problema, capturar requerimientos de forma precisa y formar una fundamentación para la implementación.

Inductivo deductivo

El estudio de los elementos relacionados con el campo de acción permitió deducir los elementos que iba a comprender la estrategia de seguridad e inducir las posibles aplicaciones del mismo en proyectos con características similares.

Analítico sintético

El análisis de los mecanismos de seguridad para cada uno de los subsistemas del almacén de datos, permitió comprobar cómo la comunicación segura entre estos conlleva a lograr mayor seguridad en las soluciones de almacenes de datos.

Métodos empíricos: los métodos empíricos utilizados fueron el Análisis documental y la Entrevista.

Análisis documental

Al revisar la literatura especializada relacionada con el tema, así como la extracción de los referentes teóricos que sustentan la investigación.

Entrevista

Las entrevistas se realizaron de manera planificada en los laboratorios de producción que desarrollan almacenes de datos del centro DATEC. La realización de las mismas permitió caracterizar el estado actual de las experiencias en el desarrollo de este tipo de soluciones.

Estructura del trabajo

Para el cumplimiento de los elementos planteados, la investigación se estructuró en: introducción, tres capítulos, conclusiones, recomendaciones, referencias bibliográficas, glosario y anexos.

Capítulo 1: presenta un estudio del estado del arte de las arquitecturas de los almacenes de datos, de los factores de riesgos y vulnerabilidades que hacen a estas soluciones particularmente susceptible a ataques, del proceso de implementación de mecanismos de seguridad, de principios

y estrategias de seguridad para soluciones de AD, de estándares y herramientas relacionadas; caracterizando y analizando desde una posición crítica los elementos estudiados.

Capítulo 2: presenta la estrategia de seguridad que se propone, describiendo sus componentes fundamentales: acciones a seguir, principios de seguridad, mecanismos de seguridad y una propuesta de arquitectura que soporta la implementación de mecanismos de seguridad en cada subsistema del AD. Además, se expone una propuesta para realizar modificaciones a la herramienta de visualización de datos Pentaho BI Server 6.1 utilizada en el centro DATEC de la UCI; debido a las limitaciones relacionadas con la seguridad en su configuración básica.

Capítulo 3: se presentan algunos de los resultados obtenidos al aplicar la estrategia de seguridad en el AD "Sistema de análisis estadísticos para los procesos eleccionarios en Cuba" desarrollado por el centro DATEC de la UCI para la Comisión Electoral Nacional (CEN). Además, se muestran y se analizan los resultados de validación de la estrategia de seguridad a través de la aplicación del método Delphi, el cálculo del coeficiente de Kendall para realizar pruebas de concordancia y la utilización de técnicas multicriterios para evaluar el índice de impacto al aplicar la estrategia de seguridad propuesta.

CAPÍTULO I. MARCO TEÓRICO REFERENCIAL DEL PROCESO DE IMPLEMENTACIÓN DE MECANISMOS DE SEGURIDAD EN EL DESARROLLO DE ALMACENES DE DATOS

En este capítulo se expone un estudio del estado del arte de las principales arquitecturas de los almacenes de datos, de los factores de riesgos y vulnerabilidades que hacen a estas soluciones particularmente susceptibles a ataques. Se realiza una valoración del proceso de implementación de mecanismos de seguridad y de principios de seguridad necesarios para diseñar cualquier medida de protección de la información. Se realiza además un diagnóstico acerca de las estrategias de seguridad para soluciones de AD del centro DATEC de la UCI, de estándares y herramientas relacionadas; caracterizando y analizando desde una posición crítica los elementos estudiados.

1.1 Almacenes de datos. Arquitectura

Con el objetivo de proveer un soporte que permitiera extraer conocimiento de la información almacenada por las empresas, a principio de la década de los años 90, Bill Inmon introdujo el tema de almacenes de datos. Los definió como una colección de datos orientado a las materias, integrado, no volátil, que varía con el tiempo y que sirve de soporte al proceso de toma de decisiones (Trujillo, y otros, 2013).

Una de las razones por las cuales las soluciones de AD se desarrollan rápidamente es debido a que presentan una arquitectura muy entendible. En la Figura 1, se representan los subsistemas fundamentales que están presentes en su arquitectura general, según Ralph Kimball (Kimball, y otros, 2004).

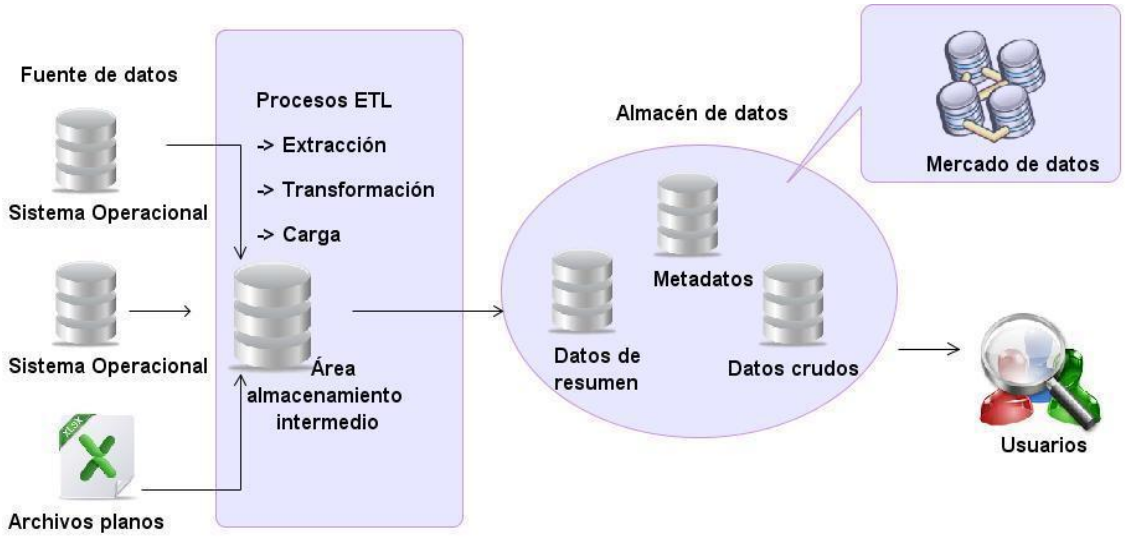


Figura 1: Arquitectura general propuesta por Kimball. (Kimball, y otros, 2004).

Las fuentes de datos que componen el subsistema Fuente de datos pueden ser: bases de datos operacionales de la organización, base de datos privadas, bases de datos públicas, ficheros de texto plano, entre otros. En el subsistema Procesos ETL es donde se realizan las actividades de extracción, transformación y carga de los datos, utilizando un área de almacenamiento intermedio conocida como "*staging area*".

Un almacén de datos es una colección de datos relacionados. Representa la unión de varios mercados de datos, los cuales almacenan la información de un área de negocio específica de la empresa. Los AD incluyen metadatos, datos de resumen, datos crudos y datos de minería (Kimball, y otros, 2004). Kimball plantea que el desarrollo de los procesos ETL representa el 70% de las tareas en la construcción de un AD; para lo cual se deben tener en cuenta dos flujos simultáneos: Planeación - Diseño, y Flujo de datos. La Figura 2 y Figura 3 muestran las etapas que sigue cada flujo respectivamente.

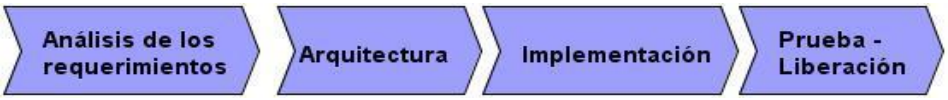


Figura 2: Hilo de Planeación – Diseño. (Kimball, y otros, 2004).

Los elementos iniciales a tener en cuenta en este proceso son los requerimientos, y dentro de estos se destacan los requisitos de seguridad. Posteriormente, se deben considerar en el diseño de la Arquitectura y deben ser la guía para las etapas posteriores de Implementación y Prueba - Liberación.

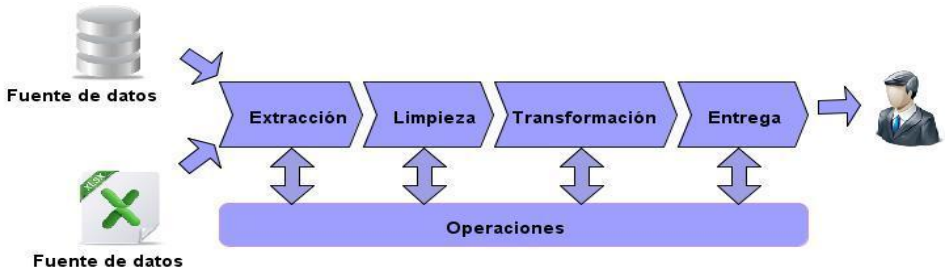


Figura 3: Flujo de datos en el proceso ETL. (Kimball, y otros, 2004).

Con la finalidad de facilitar y gestionar el funcionamiento de los procesos de un AD, existen herramientas especializadas entre las que se encuentran las herramientas de ETL, que apoyan al proceso de flujo de datos representado en la Figura 3. La herramienta utilizada en los proyectos del centro DATEC es la herramienta Pentaho Data Integration 6.1 (Bouman, y otros, 2009).

Inmon propuso en el 2008 una arquitectura hacia una nueva generación de los AD conocida como "*Data Warehouse* 2.0", la cual consiste en una evolución del modelo tradicional. En la Figura 4 se representan los sectores o subsistemas que destaca la arquitectura propuesta, estos son: **Interactivo**, **Integrado**, **Línea Cerca** y **Archivado** (Inmon, y otros, 2008).

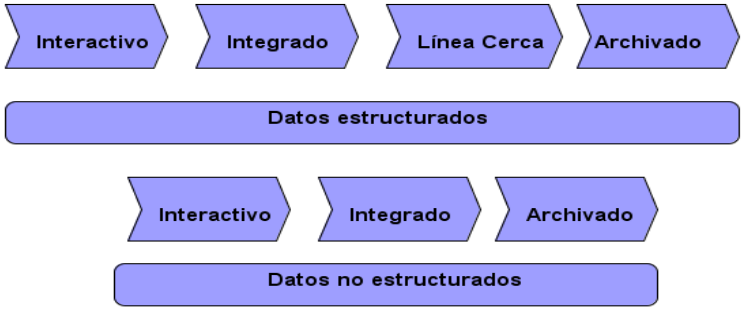


Figura 4: Arquitectura general propuesta por Inmon. (Inmon, y otros, 2008).

Los datos se almacenan en el sector **Interactivo** de dos maneras: mediante otra aplicación de almacenamiento de datos fuera del AD, que captura la información como un subproducto de una transacción (datos estructurados), o mediante procesos de ETL. En el primer caso los datos se

trasladan directamente al sector **Interactivo** y en el último caso, la aplicación ejecuta la transacción y se envían los datos a través de los procesos ETL al sector **Interactivo**. La carga de trabajo que transita a través del sector **Interactivo** es pequeña y rápida.

En el flujo de procesamiento, los datos del sector **Interactivo** se trasladan al sector **Integrado**, transformándose en datos corporativos. También pueden llegar directos al sector **Integrado** si son datos no estructurados. El tiempo de respuesta en el sector **Integrado** varía de segundos a horas. Normalmente existe un gran volumen de datos y por lo general contiene información de un período de 3 a 5 años. El sector **Línea Cerca** es un espejo o imagen del sector **Integrado**, es opcional pues generalmente se introduce en el flujo de los datos cuando existe un gran volumen de información y se trasladan al mismo los datos con una baja probabilidad de acceso. Finalmente en el flujo, los datos pasan al sector **Archivado** cuando la probabilidad de acceso se hace muy pequeña y la información tiene un período de más de 5 años.

Esta arquitectura se ha convertido en un paradigma para almacenes de datos modernos. En (Data Warehousing 2.0 Modeling and Metadata Strategies for Next Generation Architectures, 2010) se discutió su evolución y se describió la necesidad de los metadatos para controlar las actividades en cada uno de sus sectores. De igual manera en (Data Warehousing 2.0 and SQL Server: Architecture and Vision, 2009) Bill Inmon destaca este paradigma unido con SQL Server para el procesamiento paralelo de grandes volúmenes de datos.

Ralph Kimball y Bill Inmon son dos autores muy importantes en el desarrollo conceptual de un almacén de datos. Las concepciones que plantean tienen el mismo objetivo, sin embargo, su forma de obtenerlo difiere bastante. La arquitectura propuesta por Inmon y su visión se consideran como una aproximación "*top-down*" o descendente del problema, su intención es desarrollar el almacén de datos y luego desarrollar los mercados de datos. La arquitectura y visión de Kimball es conocida como una aproximación "*bottom-up*" o ascendente del problema, al construir los AD a través de los mercados de datos (Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models, 2004).

Debido al crecimiento en el volumen de los datos, el desarrollo de las nuevas tecnologías y aplicaciones, nuevas necesidades de análisis, el auge en el desarrollo de los AD y otras razones; los desarrolladores han aprendido a fusionar las tendencias propuestas por Kimball y Bill Inmon e implementar lo mejor de ambas percepciones. Particularmente, en las soluciones de AD desarrolladas por el centro DATEC de la UCI, se sigue la visión "*bottom-up*" al utilizar como metodología de desarrollo una adaptación del ciclo de vida propuesto por Kimball (González Hernández, 2013).

1.2 Almacenes de Datos. Factores de riesgos y vulnerabilidades

Las soluciones de almacenes de datos se centran en los datos valiosos de los diferentes tipos de empresas, tales como: bancos, oficinas estadísticas, compañías, instituciones, organizaciones, entre otros. En estas es un requisito esencial preservar la integridad, confidencialidad y disponibilidad de la información, sin embargo, los datos están en riesgo durante cada una de las fases del desarrollo de un almacén de datos.

La Agencia Europea de Redes y Seguridad de la Información (ENISA) es un centro de experiencia en la seguridad de la red y la información de la unión europea. Anualmente esta organización publica un informe que muestra una recopilación de amenazas y una evaluación del impacto de

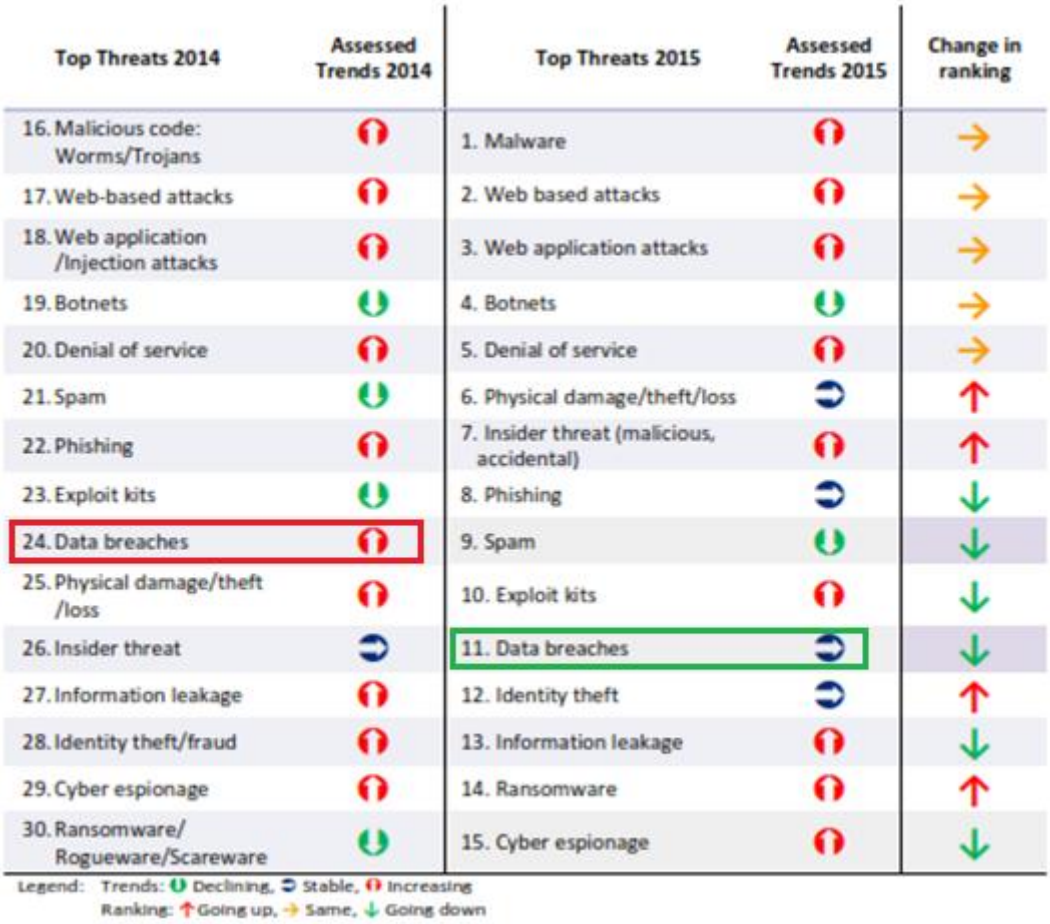
las mismas. Los datos analizados de los meses: diciembre del año 2013 y diciembre del año 2014, demostraron un incremento de aproximadamente el 25% en el tipo de amenaza: violación de los datos; respecto al informe realizado en el año 2013 (Véase, Figura 5). En el período que se examinó se identificaron muchos tipos de violaciones a los datos. La mayoría tuvieron un impacto significativo para las empresas, afectaron gran cantidad de información de los clientes y partes interesadas. De acuerdo con informes extensos de violación de datos, las causas de esta amenaza son las siguientes: contraseñas débiles, redes y aplicaciones vulnerables, programas malignos, incorrecta autenticación de usuarios, otras amenazas internas, manipulación y errores en bases de datos, entre otras. (ENISA, 2016) .

Top Threats 2013	Assessed Trends 2013	Top Threats 2014	Assessed Trends 2014	Change in ranking
1. Drive-by downloads (renamed to Web-based attacks)	↑	1. Malicious code: Worms/Trojans	↑	↑
2. Worms/Trojans	↑	2. Web-based attacks	↑	↓
3. Code Injection	↑	3. Web application /Injection attacks	↑	→
4. Exploit Kits	↑	4. Botnets	↓	↑
5. Botnets	→	5. Denial of service	↑	↑
6. Physical Damage/Theft/Loss	↑	6. Spam	↓	↑
7. Identify Theft/Fraud	↑	7. Phishing	↑	↑
8. Denial of Service	↑	8. Exploit kits	↓	↓
9. Phishing	↑	9. Data breaches	↑	↑
10. Spam	→	10. Physical damage/theft /loss	↑	↓
11. Rogueware/Ransomware / Scareware	↑	11. Insider threat	→	(NA. new threat)
12. Data Breaches	↑	12. Information leakage	↑	↑
13. Information Leakage	↑	13. Identity theft/fraud	↑	↓
14. Targeted Attacks (renamed to Cyber espionage, merged with Watering Hole)	↑	14. Cyber espionage	↑	→
15. Watering Hole (threat consolidated with other threats/attack vector)	↑	15. Ransomware/Rogueware/ Scareware	↓	↓

Legend: Trends: ↓ Declining, → Stable, ↑ Increasing
Ranking: ↑ Going up, → Same, ↓ Going down

Figura 5: Descripción y comparación de amenazas período (2013-2014). (ENISA, 2016).

No solo los datos personales están expuestos a amenazas, la información valiosa para las instituciones es considerada objetivo prioritario para ataques informáticos. A diferencia del año 2014 que fue considerado el "año de la violación de los datos", en el año 2015 este tipo de violación se mantuvo estable (ligero descenso) como se muestra en la Figura 6. Este es un resultado positivo en general, pero la violación de los datos continúa siendo objeto de análisis. Las amenazas relacionadas son las siguientes: malware, daños físicos / robo / pérdida, ataques basados en la web, ataques de aplicaciones web, suplantación de identidad, correos no deseados, amenaza de información privilegiada, fuga de información y robo de identidad (ENISA, 2016).



Top Threats 2014	Assessed Trends 2014	Top Threats 2015	Assessed Trends 2015	Change in ranking
16. Malicious code: Worms/Trojans	↑	1. Malware	↑	→
17. Web-based attacks	↑	2. Web based attacks	↑	→
18. Web application /Injection attacks	↑	3. Web application attacks	↑	→
19. Botnets	↓	4. Botnets	↓	→
20. Denial of service	↑	5. Denial of service	↑	→
21. Spam	↓	6. Physical damage/theft/loss	⊃	↑
22. Phishing	↑	7. Insider threat (malicious, accidental)	↑	↑
23. Exploit kits	↓	8. Phishing	⊃	↓
24. Data breaches	↑	9. Spam	↓	↓
25. Physical damage/theft /loss	↑	10. Exploit kits	↑	↓
26. Insider threat	⊃	11. Data breaches	⊃	↓
27. Information leakage	↑	12. Identity theft	⊃	↑
28. Identity theft/fraud	↑	13. Information leakage	↑	↓
29. Cyber espionage	↑	14. Ransomware	↑	↑
30. Ransomware/ Rogueware/Scareware	↓	15. Cyber espionage	↑	↓

Legend: Trends: ↓ Declining, ⊃ Stable, ↑ Increasing
Ranking: ↑ Going up, → Same, ↓ Going down

Figura 6: Descripción y comparación de amenazas período (2014-2015). (ENISA, 2016).

Particularmente, en la literatura existen factores de riesgos que hacen a los almacenes de datos susceptibles a ataques, como se menciona en (A Schematic Technique Using Data type Preserving Encryption to Boost Data Warehouse Security, 2011):

- Los datos extraídos se trasmiten a través de líneas de comunicación inseguras.
- Los datos extraídos se almacenan en una variedad de sistemas informáticos y medios extraíbles.
- El proceso de extracción produce archivos intermedios, los cuales cargan información sensible desde la fuente de datos.
- El mantenimiento de los atributos de seguridad para las tablas del AD es extremadamente consumidor de tiempo ante el constante cambio organizacional.
- Los usuarios a menudo recuperan información de los datos del AD y crean mercados de datos, dando lugar a una amplia distribución de información sensible.
- La práctica de la seguridad es con frecuencia descuidada, debido al cronograma apretado de los proyectos.

Asociado a estos factores de riesgos se identifican las siguientes vulnerabilidades en las soluciones de almacenes de datos:

- Líneas de comunicación inseguras.
- Medios de almacenamientos inseguros.
- Datos sensibles mal gestionados.
- Manipulación incorrecta y errores de bases de datos.

- Exposición de las copias de respaldo de la información.
- Auditorías débiles en bases de datos.
- Configuración errónea de la seguridad.
- Contraseñas débiles.
- Incorrecta autenticación de los usuarios.
- Privilegios excesivos e inutilizados.
- Características de bases de datos innecesariamente habilitadas.
- Bases de datos sin actualizar.
- Redes y aplicaciones vulnerables.
- Aplicaciones web vulnerables teniendo en cuenta los 10 tipos de vulnerabilidades más comunes para estas soluciones, definidas por el Proyecto abierto de seguridad de aplicaciones web (OWASP por sus siglas en inglés: Open Web Application Security Project) (González Brito, 2016).

Los cambios constantes en las peticiones de los usuarios y en las fuentes de datos precisan no solo a ser más flexibles, sino también a controlar la seguridad de la información de manera más eficaz. Un aspecto muy importante a considerar es que la información no debe ser tratada de forma estática, pues los datos almacenados en los AD son relativos a un periodo de tiempo y deben ser incrementados periódicamente. En gran medida, el éxito de las organizaciones depende de la correcta gestión, seguridad y confidencialidad de la información (A UML 2.0/OCL Extension for Designing Secure Data Wharehouses, 2006). En (Data Warehousing - Security, 2016) se resumen las actividades que pudieran ser afectadas por razones de seguridad en los AD, de la siguiente manera:

- Acceso de los usuarios.
- Carga de datos.
- Traslado de los datos.
- La generación de consultas.
- Otras.

1.3 Almacenes de Datos. Requisitos de Seguridad

Existen muchas definiciones relacionadas con la seguridad, estas son resumidas por las siglas CIA, acrónimo de confidencialidad, integridad y disponibilidad. En (Basic concepts and taxonomy of dependable and secure computing., 2004) se define la confidencialidad como la ausencia de la divulgación no autorizada de la información, la integridad como la ausencia de alteraciones del sistema y disponibilidad como la correcta disposición de los servicios del sistema.

Los requisitos de seguridad para los AD son similares a los de otros sistemas informáticos distribuidos. Implementar mecanismos de control interno para garantizar la confidencialidad, integridad y disponibilidad de los datos en un entorno distribuido es de suma importancia. Según (Cryptome, 2013) la confidencialidad significa que la información debe permanecer en secreto y solo las personas autorizadas para acceder a ella pueden recibir acceso. La integridad requiere la protección contra la modificación intencionada o accidental, se refiere no solo a la integridad de la información sino también a la integridad del origen, es decir, a la integridad de la fuente de información. La disponibilidad es la característica que asegura la disposición de datos a los usuarios autorizados cuando los necesiten.

Las vulnerabilidades relacionadas con la seguridad pueden afectar el rendimiento del almacén de datos. Por lo tanto, es importante determinar los requisitos de seguridad en las primeras etapas de desarrollo. Durante la fase de diseño del almacén de datos, pueden presentarse cambios en la fuente de datos y en el tipo de acceso por parte de los usuarios, por lo cual se debe analizar el impacto que esto pudiera provocar. Debido a esto, se deben tener en cuenta los siguientes aspectos durante la fase de diseño (Data Warehousing - Security, 2016):

- Si las nuevas fuentes de datos requerirán nuevas restricciones de seguridad y/o de auditoría para ser implementadas.
- Si pudieran existir cambios en el acceso por parte de los usuarios.

Esta situación se presenta cuando los futuros usuarios y las fuentes de datos no son bien conocidos. En tal escenario, se deberá utilizar el conocimiento del negocio y el objetivo del almacén de datos para conocer los requisitos que pudieran cambiar. Se sugiere, en primer lugar clasificar los datos y posteriormente clasificar los usuarios sobre la base de los datos que pueden acceder. Estas clasificaciones y otros elementos propuestos se exponen de la siguiente forma (Data Warehousing - Security, 2016):

Clasificación de datos

- Los datos pueden ser clasificados de acuerdo con su sensibilidad. Los datos altamente sensibles se clasifican como altamente restringidos y los datos menos sensibles se clasifican como menos restringidos.
- Los datos también se pueden clasificar de acuerdo a la función de trabajo. Esta restricción permite que solo determinados usuarios accedan a datos particulares, según interés y responsabilidad.

Clasificación de usuario

- Los usuarios se pueden clasificar de acuerdo con la jerarquía de los usuarios de una organización. Es decir, los usuarios pueden ser clasificados por departamentos, secciones, grupos, y así sucesivamente.
- Los usuarios también pueden clasificarse de acuerdo con su función.

Los requisitos a auditar para verificar estos elementos pueden clasificarse de la siguiente manera:

- Conexiones.
- Desconexiones.
- Acceso a los datos.
- Traslado de los datos.

Requisitos de Red

La seguridad de la red es otro requisito de seguridad importante. Por lo que se deben tener en cuenta las siguientes interrogantes:

- ¿Es necesario cifrar los datos?
- ¿Existen restricciones en las rutas de red que los datos pueden tomar?

Traslado de los datos

Existen posibles implicaciones de seguridad al trasladarse los datos. Por lo cual, se plantean interrogantes como:

- ¿Dónde se almacena la fuente de datos?
- ¿Quién tiene acceso?
- ¿Están cifrados o no las versiones de copia de seguridad?
- ¿Es necesario ubicar por separado las copias de seguridad?

1.4 Propuestas de seguridad existentes para almacenes de datos

En (Kimberly, 2013) se describe la necesidad de proteger los almacenes de datos, pues constituyen el motor de fuerza de trabajo para la toma de decisiones de negocio, una tarea cada vez más desafiante con 2,5 trillones de bytes de datos creados cada día. El costo promedio de los incidentes relacionados con la seguridad en esta era de grandes volúmenes de datos ("*big data*"[4]) es aproximadamente de más de 40 millones de dólares (Barranco Fragoso, 2012). Por lo que no se puede ignorar la seguridad de datos como requisito primordial. No sólo las violaciones de datos suponen una pérdida económica, sino que también pueden afectar negativamente a la organización.

Los cambios constantes en las peticiones de los usuarios y en la fuente de datos en el desarrollo de AD requieren soluciones flexibles. Aparejado con la flexibilidad se necesita controlar la seguridad de manera eficaz pues la información es la razón de estas soluciones.

En la literatura se pueden encontrar varias iniciativas para la inclusión de la seguridad en los AD [Katic, Quirchmayr, Schiefer, Stolba y Min Tjoa, 1998; Kirkgöze, Katic, Stolda y Min Tjoa, 1997; Priebe y Pernul, 2000; Rosenthal y Sciore, 2000]. Varios de ellos se centran en principios relacionados con el control de acceso, la seguridad multinivel, aplicaciones a bases de datos federadas, aplicaciones utilizando herramientas comerciales, entre otros (A UML 2.0/OCL Extension for Designing Secure Data Wharehouses, 2006).

Existen iniciativas que intentan integrar la seguridad en el modelo conceptual de los AD (Modeling security-relevant data semantics, 1991), el filtrado y cifrado de los datos antes de su almacenamiento en los AD (View Security as the Basic for DW Security, 2000), la seguridad en la semántica del modelo de datos, y la seguridad en el modelamiento de objeto multinivel (Access control and audit model for the multidimensional modeling of data warehouses, 2006). En la literatura se encuentran además otras propuestas de seguridad como: construcción de un esquema en estrella seguro para soluciones de almacenes de datos (Building a secure star schema in data warehouse by an extension of rela-tional package from CWM, 2008), la aplicación de un enfoque basado en la arquitectura dirigida por modelos (MDA) para desarrollar almacenes de datos seguros (Applying an MDA-based approach to consider security rules in the development of secure DWs, 2009), un enfoque híbrido para asegurar los almacenes de datos, el cual integra modelos clásicos de seguridad y la aplicación del filtrado de los datos (An Improved Security Framework for Data Warehouse: A Hybrid Approach, 2010), entre otras propuestas.

4 Grandes volúmenes de datos (estructurados, no estructurados y semi-estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a una base de datos relacional para su análisis.

En (Gosaina, y otros, 2015) se realiza un análisis de estas y otras propuestas. De las 55 analizadas, 19 se centran en aspectos de seguridad de un AD. Las propuestas se compararon sobre la base de los siguientes parámetros de seguridad:

- Datos cifrados (ED): los datos contenidos en el almacén de datos son encriptados o no.
- Control de auditoría (AC): inclusión de parámetros de auditoría de seguridad en el almacén de datos.
- Extensibilidad (EX): extensibilidad del modelo del AD a nuevos requisitos de seguridad.
- Seguridad computacional independientemente del modelo (CI).
- Seguridad del modelo independientemente de la plataforma (PI).
- Seguridad del modelo dependiente de la plataforma (PS).
- Apoyo QVT: si la propuesta de seguridad brinda soporte a los procedimientos, vistas y transformaciones de consultas (QVT).
- Implementación desarrollada (I).
- Integración de datos multiplataforma: responde a la interrogante ¿el desarrollo de la tecnología de seguridad permite la integración de datos de fuentes heterogéneas? (ID).

A continuación, en la Figura 7 se observa el análisis gráfico de las consideraciones de seguridad por las propuestas en forma consolidada.

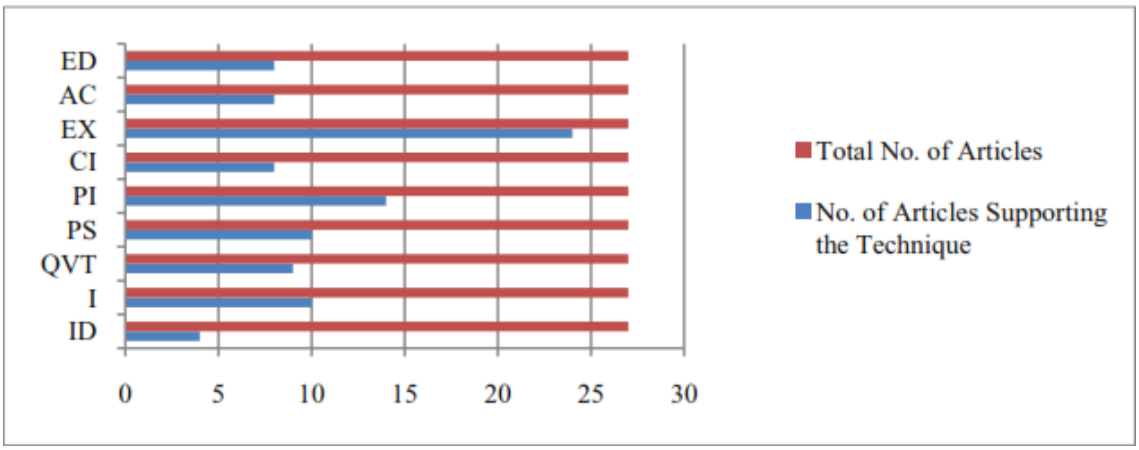


Figura 7: Número de propuestas que soportan los diferentes parámetros de seguridad. (Gosaina, y otros, 2015).

La mayoría de las propuestas soportan la extensibilidad, pero menos de la mitad brindan apoyo al control de auditoría y a la seguridad computacional independientemente del modelo. La mayor parte de las propuestas son compatibles con la extensibilidad y la seguridad independiente de la plataforma. Como se representa en la Figura 8, el 32% de las propuestas pertenecen a las áreas de OLAP y UML, mientras que otras categorías conforman el 68%. De este 68%, el 21% está relacionado con técnicas del lenguaje de modelado UML y el cifrado de datos. El 19% de las propuestas se relacionan con el área MDA, un 3% trata el cifrado de datos y los metadatos.

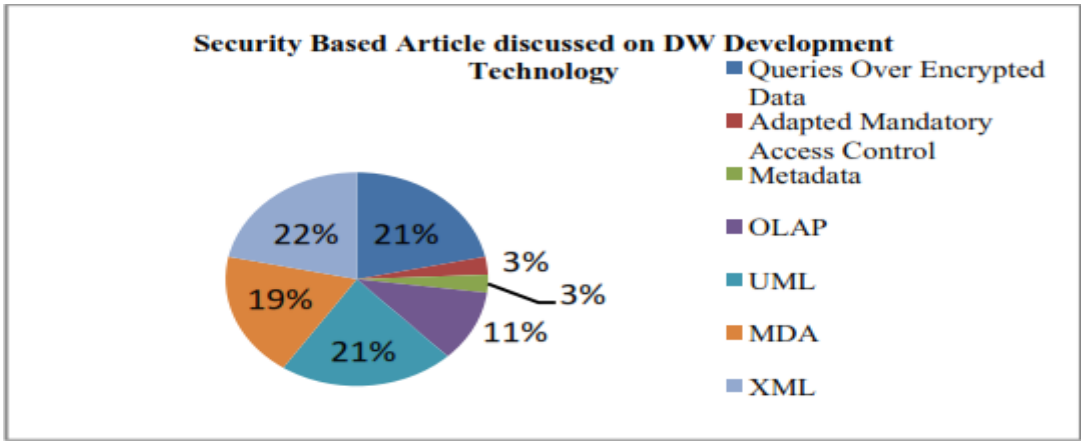


Figura 8: Porcentaje de propuestas por áreas que tratan la seguridad en el desarrollo de almacenes de datos. (Gosaina, y otros, 2015).

Estas propuestas de seguridad para AD se centran en aspectos específicos y no consideran elementos de seguridad que deben incorporarse en todas las etapas del ciclo de desarrollo.

1.5 Principios y mecanismos de seguridad

En (Seguridad Informática Conceptos generales, 2002) se definen los principios básicos considerados como el pilar de la seguridad de la información. Estos principios son:

- **Principio de menor privilegio:** establece que cada usuario debe tener los privilegios estrictamente necesarios de acuerdo a la función que desempeña dentro de la organización.
- **Seguridad no equivale a oscuridad:** se establece la percepción de que el no conocer las vulnerabilidades del sistema no lo hace inmune a fallos o amenazas, por el contrario el conocimiento de ellas conllevará a proveer las medidas de seguridad necesarias para subsanarlos.
- **Principio del eslabón más débil:** propone proteger los posibles puntos que representen una mayor vulnerabilidad para el sistema ante amenazas. Siempre haciendo énfasis en seguir las prácticas de seguridad establecidas.
- **Defensa en profundidad:** plantea el establecimiento de mecanismos de protección en varios niveles sucesivos con el objetivo de implantar múltiples defensas de la información ante cualquier amenaza que se pueda presentar.
- **Punto de control centralizado:** propone en cierto sentido figurativo y práctico una especie de embudo invertido de seguridad. Es decir, que el sistema tenga un solo punto de acceso con varios mecanismos de seguridad para acceder a la información.
- **Seguridad en caso de fallo:** expone que en caso de que una amenaza (externa o interna) tenga éxito al penetrar el sistema, el mismo debe permanecer en un estado de bloqueo de seguridad hasta tanto se solvente la situación.
- **Participación universal:** se necesita el acuerdo entre todos los usuarios para poner en práctica las medidas de seguridad planteadas y establecidas dentro de la organización con observancia constante de las mismas.
- **Principio de simplicidad:** lo que se persigue con este principio es el fácil entendimiento y manejo de la información, evitando la complejidad lo necesariamente posible; esto evitará fallos y puntos débiles.

Kimball e Inmon proponen algunos principios de seguridad para sus propuestas de arquitectura. Teniendo en cuenta que la información almacenada en un AD puede verse comprometida por disímiles ataques, Kimball propone un grupo de acciones para hacer menos vulnerable la información (Ralph, y otros, 2010). Estas son:

- **Arquitectura distribuida:** el AD empresarial debe estar compuesto por varios equipos, sistemas operativos, tecnologías de bases de datos, aplicaciones analíticas, vías de comunicación, lugares, y copias de los datos. Este mecanismo reduce en gran medida la vulnerabilidad ante ataques y fallos de un solo punto.
- **Ampliar las redes de área de almacenamiento (SAN):** una SAN es un grupo de unidades de alto rendimiento de disco y dispositivos de copias de seguridad, conectados a través de una tecnología de canal de fibra de muy alta velocidad. Esto permite conexiones desde lugares distantes, copias de seguridad entre discos a grandes velocidades, el acceso a los datos en paralelo desde múltiples sistemas de aplicación, entre otros.
- **Copias diarias hacia medios seguros:** se recomienda realizar copias de seguridad de la información hacia medios seguros, sin embargo, hay que tener en cuenta cuán difícil sería leer los datos almacenados en períodos de un año, cinco e incluso 10 años.
- **Situar estratégicamente gateway[5] de filtrado de paquetes:** se trata de aislar a los servidores claves del AD detrás de un *gateway* de filtrado de paquetes, de modo que no sea directamente accedido. El servidor de bases de datos puede recibir paquetes desde el exterior si proceden del servidor de aplicaciones.
- **Cuello de botella de la autenticación y el acceso:** se recomienda utilizar el Protocolo Ligero de Acceso a Directorios (LDAP por sus siglas en inglés: Lightweight Directory Access Protocol), con el objetivo de controlar todos los accesos fuera del *gateway* al AD. El servidor LDAP permite que los usuarios se autentiquen de manera uniforme. Una vez que el usuario se ha autenticado, el servidor de aplicación toma la decisión de mostrar la información, solo si es un usuario válido en el sistema.

Un AD está compuesto por varios subsistemas que deben ser considerados para lograr un nivel de seguridad adecuado. En (Security of Data Warehousing Server, 2010) se proponen los siguientes principios de seguridad para este tipo de soluciones: integridad y validación de los datos, enmascaramiento de datos y conservación de la privacidad, políticas de acceso y restricciones de datos, clasificación de los datos; basados todos estos principios en el ambiente del sistema (redes, servidores, sistemas operativos, aplicaciones).

- **Integridad y validación de los datos**: necesidad de asegurar que los datos cargados al sistema de almacenamiento son válidos y correctos. Incluye las actividades que deben realizarse cuando la información proviene de múltiples fuentes.

[5] Es una puerta de enlace, un nodo en una red informática que sirve de punto de acceso a otra red.

- **Enmascaramiento de datos y conservación de la privacidad**: necesidad de asegurar la privacidad y confidencialidad de los datos. Solo se hace disponible un adecuado nivel de detalle de la información.
- **Políticas de accesos y restricciones de datos**: se trata de llevar a cabo la protección de los datos con limitaciones de accesos. Las políticas de acceso y las restricciones en los datos son la base para la aplicación de métodos de auditoría.
- **Clasificación de los datos**: comprende la naturaleza de los datos almacenados y su clasificación correcta. Se indica el nivel de sensibilidad de los datos que es la base para la aplicación de todos los mecanismos de seguridad.

Inmon plantea que existen mecanismos de seguridad que pueden implementarse para evitar que las personas tengan acceso no autorizado a la información, tales como (Inmon, y otros, 2008):

- El cifrado de los datos no se encuentra dentro de la protección al acceso de la información, sino que se realiza con el objetivo de preservar su integridad. Plantea que se deben utilizar algoritmos de cifrado para verificar que la información no sea modificada y alterada por elementos maliciosos. Posteriormente, en el acceso a la información se necesita el conocimiento del algoritmo utilizado para poder descifrar los datos. Inmon considera que el cifrado de grandes volúmenes de datos es poco efectivo. Si solo se cifra una parte de los datos, se reducen sus desventajas, respecto al esfuerzo que se necesita para descifrar. Es por ello que recomienda que el cifrado de los datos se realice sobre los datos del sector **Integrado** que propone su arquitectura "*Data Warehouse 2.0*" y de forma moderada, teniendo en cuenta el rendimiento del sistema.
- Otra medida de seguridad que menciona es el uso de *firewall*[6] cuando existen conexiones desde internet a los sistemas de la empresa. Este solo controla el acceso del sector **Interactivo**, pues es donde se realiza el procesamiento activo de transacciones. El sector **Integrado**, el sector **Línea Cerca** y el sector **Archivado**, no deben ser accedidos directamente desde internet. De existir la necesidad de acceder a estos últimos, se necesita acceso autorizado desde la propia red interna de la empresa. Debido a que el sector **Línea Cerca** es una extensión del sector **Integrado**, pudieran aplicarse las mismas medidas de seguridad a ambos sectores.
- Es importante llevar a cabo tareas de monitorización, para poder hacer un seguimiento del acceso a la información. Esto permite analizar si han existido transacciones no autorizadas mediante el análisis de las trazas registradas en los *logs*.
- Por otro lado, si muchas peticiones ocurren sucesivamente y tratan de atravesar el *firewall*, el sistema debe prever que un ataque pudiera ocurrir. Si se intenta acceder con muchas contraseñas no autorizadas, el sistema debe ser sensible a ataques malintencionados y detenerse hasta que este incidente termine.

[6] Es un filtro que controla las comunicaciones que pasan a través de la red.

En (Bouman, y otros, 2009), se hace referencia a dos de estos mecanismos: el acceso a la información y la encriptación de los datos. Por otra parte, Inmon en (Inmon, y otros, 2001) plantea que los niveles de seguridad se implementan a través de diferentes enfoques tecnológicos (Véase, Figura 9) y que cada uno tiene ventajas y desventajas como se describe a continuación:

- **Seguridad basada en *firewall*:** entre las ventajas se destacan que es barata y fácil de construir, sin embargo, no protege los datos a medida que pasan a través de la red, mientras entran y salen de la base de datos. No se ofrece ninguna protección si el *firewall* es violado.
- **Inicio de sesión / cierre de sesión de seguridad:** con frecuencia se encuentra incorporado al software, es fácil de implementar y barato. Entre sus desventajas se encuentran: no protege los datos a medida que pasan a través de la red, mientras entran o salen de la base de datos y no ofrece protección cuando se incumple el inicio o cierre de sesión si son datos cifrados.
- **Vistas basadas en la seguridad del sistema gestor de base de datos (DBMS):** se encuentra incorporado al software DBMS, es barato y engorroso de implementar. No protege los datos a medida que pasan a través de la red, mientras entran o salen de la base de datos y no ofrece ninguna protección cuando las vistas han sido incumplidas.
- **Cifrado / Descifrado:** es relativamente caro, no necesariamente difícil de implementar. Entre sus ventajas se destacan la protección de los datos a medida que pasan a través de la red, mientras entran o salen de la base de datos y ofrecen protección cuando el *firewall*, el inicio / cierre de sesión y las vistas han sido incumplidos.

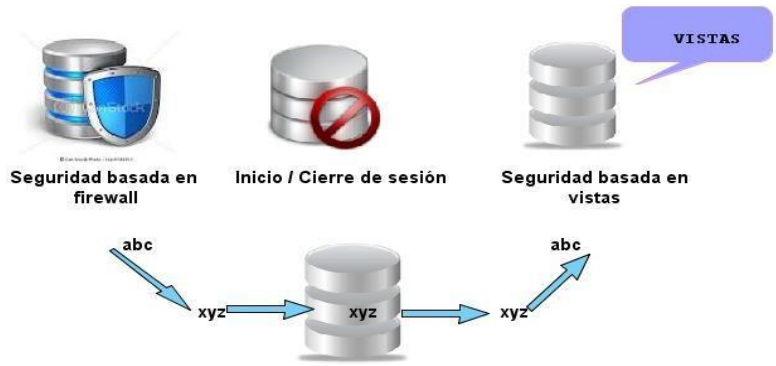


Figura 9: Mecanismos de seguridad. Inmon (Inmon, y otros, 2001).

Inmon señala que los diferentes tipos de seguridad tienen diferentes niveles de eficacia y costos, donde el nivel de eficacia es directamente proporcional a los costos y restricciones que acompañan a estos mecanismos de seguridad. Señala además que los diferentes enfoques de seguridad se pueden utilizar solos o una combinación de ellos.

En una solución de AD se debe tener en cuenta la seguridad de extremo a extremo. El entorno de este tipo de soluciones no solo está compuesto por una base de datos, por lo que se necesitan asegurar los datos durante todo el flujo de la información. Existen AD en la actualidad, en los que las bases de datos presentan riesgos de seguridad, y al mismo tiempo los archivos planos que se utilizan para cargar la información se almacenan en una ubicación no segura. Esto es un ejemplo

de los fallos de seguridad que pueden surgir cuando el proceso de desarrollo del almacén de datos no ha sido diseñado pensando en la seguridad (Security and the Data Warehouse, 2005).

1.6 Estado de la seguridad en soluciones de AD del centro DATEC

La metodología de desarrollo utilizada para la construcción de los almacenes de datos en el centro DATEC especifica las siguientes etapas en las que deben implementarse los requisitos de seguridad (González Hernández, 2013):

- En las actividades de administración de la seguridad en el subsistema de visualización.
- En la definición de la arquitectura con la creación de la vista de seguridad se definen los diferentes niveles de protección que deberá poseer el sistema (como parte de la Vista de Tecnología).
- En el artefacto Especificación del modelo dimensional donde se deben describir las políticas de seguridad, indexado y particionado de los datos. Además, se debe definir la estrategia de respaldo y recuperación de la información almacenada.
- En el artefacto Especificación de la aplicación para el usuario final, se definen todas las características del subsistema de Presentación de información desde el diseño de los cubos OLAP[7] hasta la implementación de las consultas MDX[8] y las políticas de seguridad definidas en la capa de visualización.
- En las actividades del diseño del subsistema de almacenamiento, en la fase de diseño e implementación, se definen los esquemas de la base de datos y la estrategia de seguridad para el acceso a los datos almacenados en el AD.
- En las actividades de implementación del subsistema de visualización de información, en la fase de diseño e implementación se implementan los reportes candidatos y las políticas de seguridad del subsistema de visualización de información.

De esta manera se establecen las actividades, fases que deben desarrollarse para implementar los requisitos de seguridad y los artefactos como resultado de su implementación. Inconvenientemente, en la metodología de desarrollo no se abordan los procedimientos necesarios que deben guiar dicha especificación e implementación.

Por otro lado, los AD desarrollados en el centro DATEC se centran en un grupo limitado de requisitos de seguridad. La mayoría de los proyectos se limitan a garantizar la seguridad en el diseño físico del despliegue de la solución, sin tener en cuenta la seguridad necesaria en los diferentes subsistemas presentes en su desarrollo: Fuente de datos, Procesos ETL y Visualización de los datos.

La arquitectura utilizada en los proyectos no agrupa la información almacenada en el AD, en relación al valor de los datos y la gravedad de las amenazas a las que puedan estar sometidos. Esto imposibilita que se puedan definir diferentes mecanismos de seguridad, teniendo en cuenta dicha relación.

[7] Por sus siglas en inglés de On-Line Analytical Processing.

[8] Por sus siglas en inglés de Multi-Dimensional expressions.

La herramienta Pentaho BI Server 6.1 utilizada para la visualización y el análisis de la información brinda funcionalidades limitadas respecto a la seguridad, debido a que en su configuración básica: los mecanismos de autenticación se especifican de forma manual editando directamente los ficheros de configuración, las contraseñas de los usuarios no se validan eficazmente y no existe la posibilidad de restringir el acceso mediante direcciones IP.

Como se evidencia, no se encontró una estrategia de seguridad para el desarrollo del ciclo de vida de estas soluciones, que incluya: acciones a seguir, propuestas de principios y mecanismos de seguridad, una arquitectura enfocada a la seguridad y la utilización de herramientas de desarrollo que solucionen los principales problemas de seguridad de la información.

1.7 Estándar ISO/IEC 27001

En la actualidad existen estándares, regulaciones, modelos y recomendaciones relacionados con la gestión de la calidad de la información. La Organización Internacional para la Estandarización (ISO) y la Comisión Electrónica Internacional (IEC) han desarrollado estándares que constituyen una referencia a nivel internacional para el logro de la seguridad de la información. Específicamente el estándar ISO/IEC 27001 (27001, 2005) ofrece un modelo para el establecimiento, implementación, operación, monitorización, revisión, mantenimiento y mejora del sistema de gestión de la seguridad de la información. Las acciones propuestas se encuentran organizadas en un círculo de acciones cíclicas: Planificar, Hacer, Verificar y Actuar (PDCA) para alcanzar los objetivos propuestos. Este estándar está enfocado a los procesos de la organización, incluye 133 controles de seguridad, divididos en 11 dominios y 39 objetivos de control. Incluye controles técnicos, otros relacionados con los recursos humanos y otros de tipo organizativo. El estándar ISO/IEC 27001 es el que tiene mayor aceptación a nivel internacional y constituye una norma certificable (Montesino Perurena, 2013).

El círculo PDCA (por sus siglas en inglés: *plan-do-check-act*) también conocido como ciclo de Control de *Deming*, círculo PHVA (por sus siglas en español) o espiral de mejora continua, desde su creación ha sido utilizado en múltiples empresas por sus sistemas de gestión de calidad (SGC) y los sistemas de gestión de la seguridad de la información (SGSI). Es muy utilizado para gestionar aspectos de calidad tales como (ISO 9000), medio ambiente (ISO 14000), salud y seguridad ocupacional (OHSAS 18000), o inocuidad alimentaria (ISO 22000) y se ha convertido en un símbolo de la mejora continua (Calidad & Gestión. Consultoría para empresas, 2015).

En la presente investigación se analizaron los controles técnicos propuestos por el estándar ISO/IEC 27001 para identificar cuáles pudieran ser aplicados al gestionar la seguridad de las soluciones de AD del centro DATEC de la UCI.

1.8 Acciones a seguir para la seguridad de los almacenes de datos

En la literatura existe un gran número de acciones a considerar para elevar la seguridad de un almacén de datos, sin embargo, las características específicas de la solución respecto a la naturaleza de los datos, el sistema de almacenamiento, la arquitectura, entre otros aspectos, definen el número de acciones a considerar. Un ejemplo de estas acciones se exponen en (A survey on current security perspectives in data warehouses, 2016), las cuales son:

Acción 1: Identificación de los datos.

La primera acción consiste en identificar todos los datos corporativos. Se trata de una acción a menudo ignorada, pero crítica para cumplir con los requisitos de seguridad del AD, ya que constituye la base para las acciones posteriores. La información recogida debe ser organizada, documentada y conservada para la siguiente acción.

Acción 2: Clasificación de los datos.

La clasificación de todos los datos es necesaria para satisfacer los requisitos de seguridad relacionados con la confidencialidad, integridad y disponibilidad de manera satisfactoria. La realización de esta acción requiere de la participación de todas las partes interesadas. Los datos se clasifican generalmente sobre la base de la criticidad o sensibilidad a la divulgación, modificación y destrucción de la información en:

- Públicos: los datos menos sensibles.
- Confidenciales: los datos moderadamente sensibles.
- Secretos: los datos más sensibles.

Independientemente de qué categorías se utilizan para clasificar los datos sobre la base de la sensibilidad; el objetivo universal de esta acción es clasificar los datos en categorías mediante el aumento de grados de sensibilidad. De modo que diferentes medidas de protección se puedan utilizar para diferentes categorías. La clasificación de datos en diferentes categorías no es una tarea fácil. Ciertos datos representan una mezcla de dos o más categorías, dependiendo del contexto utilizado, por ejemplo: el tiempo, la ubicación y las leyes vigentes.

Acción 3: Cuantificar el valor de los datos.

El proceso de cuantificación se refiere principalmente sobre la asignación de "valor" a los datos agrupados en diferentes categorías de sensibilidad. Por sí mismo, los datos no tienen valor intrínseco. El valor definitivo de los datos es a menudo medible por el costo para: reconstruir los datos perdidos, restaurar la integridad de los datos corruptos o datos interceptados, no tomar decisiones a tiempo debido a la denegación de servicio, o pagar la responsabilidad financiera para la divulgación pública de datos confidenciales, entre otros. El valor de los datos también puede incluir ingresos procedentes de la fuga de secretos comerciales a los competidores y la utilización anticipada de datos financieros antes de su publicación.

Acción 4: Identificar las vulnerabilidades de los datos y sus costos.

Esta acción requiere la identificación y documentación de las vulnerabilidades asociadas al AD. Algunas vulnerabilidades comunes están relacionadas con: la seguridad incorporada en los BMS[9], limitaciones en los DBMS, ataques de inferencia, factor de disponibilidad, factores humanos, amenazas internas, amenazas externas, factores naturales, entre otros.

Acción 5: Identificar las medidas de protección de datos y sus costos.

Se deben considerar las vulnerabilidades identificadas en la acción anterior para determinar la protección de los datos del AD. Algunas medidas de protección de los datos del AD incluyen:

[9] Por sus siglas en inglés de Data Base Management System.

gestión de recursos humanos que conozcan elementos de seguridad, clasificación del acceso de los usuarios, controles de acceso, controles de integridad, encriptación de los datos, particionado, entre otros. Los costos estimados de cada medida de seguridad deben ser determinados y documentados para la siguiente acción. La medición de los costos por lo general implica la determinación de los costos de desarrollo, implementación y mantenimiento de cada medida de seguridad.

Acción 6: Seleccionar las medidas de seguridad teniendo en cuenta la relación costo-efectividad.

Todas las medidas de seguridad implican gastos, y los gastos de seguridad requieren justificación. Esta acción se basa en los resultados de las acciones anteriores para evaluar el impacto de los datos corporativos en situación de riesgo, y optar por medidas de seguridad rentables para proteger los datos contra las vulnerabilidades conocidas.

Acción 7: Evaluar la efectividad de las medidas de seguridad.

Se pretende minimizar la ocurrencia de ataques, o estar preparados para recuperarse rápidamente ante su impacto. La institución estará preparada para llevar a cabo cualquiera de estas acciones, si evalúa la eficacia de las medidas de seguridad sobre una base continua. La evaluación de la eficacia de las medidas de seguridad debe llevarse a cabo de forma continua y debe determinar si las medidas son: sencillas y directas, analizadas cuidadosamente, probadas y verificadas, si se utilizan correctamente y de forma selectiva a fin de que no se excluyan los accesos permitidos, flexibles de modo que puedan responder eficazmente a las necesidades cambiantes de seguridad, razonablemente eficientes en términos de tiempo, espacio de memoria y que no afecten negativamente a los recursos informáticos protegidos.

Se trata de un grupo de acciones básicas para la planificación de la seguridad de los datos, aunque pudiera no ser necesaria la aplicación de todas las acciones, e incorporar otros elementos necesarios. La etapa de planeación es de vital importancia para lograr soluciones que implementen mecanismos de seguridad de forma eficaz. En (Security of Data Warehousing Server, 2010), en (Data Warehouse Security Considerations, 2016) y otros, se destacan también otro número de acciones similares. En la presente investigación se pretenden adaptar estas acciones a las soluciones de AD que son objetos de estudio.

1.9 Plataforma de Inteligencia de Negocio. Suite de Pentaho

La Suite de Pentaho proporciona un espectro completo de herramientas de inteligencia de negocio, reportes, análisis, *dashboards*, minería de datos e integración de datos. Ofrece además una serie de servicios críticos entre los que están la autentificación, programación de tareas, seguridad y servicios web. Este conjunto de herramientas y servicios forman una plataforma integral de inteligencia de negocio, convirtiendo a Pentaho en el proveedor líder de soluciones BI de código abierto (Aguilar Mayorga, y otros, 2009).

El modelo de negocio de código libre y comercial de Pentaho elimina las licencias de software, proporciona soporte, servicios y mejoras. En los últimos años los productos de Pentaho han sido descargados por más de tres millones de usuarios; proporcionando sistemas para empresas pequeñas, medianas y grandes (Pentaho BI, 2011) que permiten realizar las siguientes acciones:

- Informar: acceder a los datos y suministrar información a toda la empresa.
- Analizar: explorar y analizar los datos interactivamente y de forma muy rápida.

- Sintetizar: alcanzar visibilidad con medidas y ratios a través de cuadros de mando.
- Integrar: integrar datos de diferentes orígenes y desde múltiples fuentes.
- Investigar: descubrir patrones ocultos e indicadores de tendencias futuras.

En la Figura 10 se muestra una representación gráfica de los componentes que conforman la Suite de Pentaho. Las principales capas son claramente identificadas, con la capa de presentación en lo más alto y la capa de datos e integración de aplicaciones en lo más bajo. Todos los componentes están expuestos vía *Web Services* para facilitar la integración con Arquitecturas Orientadas a Servicios (SOA por sus siglas en inglés: Service Oriented Architecture) (Gravitar, 2016). Las áreas funcionales más importantes son: reportes, análisis, *dashboards* y administración de procesos (constituyen la capa intermedia), en tanto que la plataforma BI en sí misma contiene elementos para la seguridad y administración (Aguilar Mayorga, y otros, 2009) .

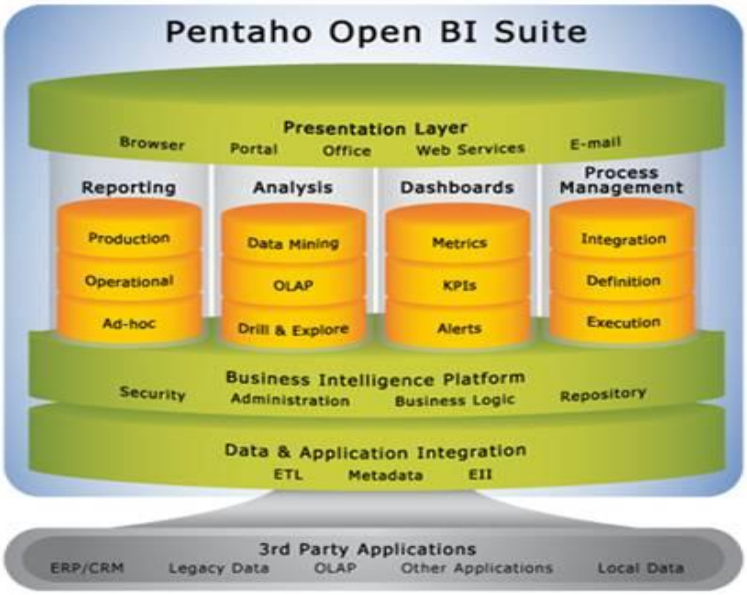


Figura 10: Arquitectura funcional de Pentaho. (Aguilar Mayorga, y otros, 2009).

Entre las herramientas de desarrollo que provee Pentaho para facilitar la toma de decisiones de las empresas, se encuentran: Pentaho Reporting, Pentaho Analisys Services, Pentaho DashBoard, Pentaho Data Integration, Pentaho Data Mining y Pentaho BI Server. En el próximo epígrafe se realizará una breve descripción de las principales características y funcionalidades que presenta la herramienta Pentaho BI Server.

1.9.1 Pentaho BI Server

Pentaho BI Server ofrece dos versiones de su solución, una versión comunitaria gratuita, orientada principalmente al mundo académico y una versión privativa, orientada a la implementación profesional. Esta herramienta funciona como un sistema de gestión basado en la web, está diseñada para integrarse fácilmente con cualquier proceso de negocio y proporciona una interfaz para la visualización de la información analizada, en forma de gráfica o agrupada por un determinado atributo que presenta el negocio (help.pentaho.com). Pentaho BI Server permite integrar diferentes mecanismos de autenticación que asisten a la seguridad de la herramienta, entre los cuales se encuentran los basados en un LDAP y la utilización del CAS. En los próximos epígrafes se presenta una breve explicación de estos mecanismos.

La plataforma Pentaho emplea "*Spring Security*[10]" para tratar la autentificación y autorización de usuarios. Esto es una solución de seguridad estándar del marco de trabajo Java Spring. Proporciona la lógica a seguir si un usuario necesita ser autenticado y puede delegar solicitudes de autenticación a un mecanismo de autenticación externo, tales como un servidor de base de datos, un directorio LDAP, o autenticación NTLM[11] sobre una red Windows (Aguilar Mayorga, y otros, 2009).

Específicamente en la versión Pentaho BI Server 6.1 utilizada en el centro DATEC de la UCI, cada permiso de operación dado es específico para cada rol como se muestra en la siguiente tabla (help.pentaho.com):

Tabla 1: Roles y permisos de Pentaho BI Server 6.1. (help.pentaho.com).

Rol	Rol por defecto	Funcionamiento de permisos por defecto
Administrador	admin	Administración de seguridad Acceso a contenido por horario Lectura de contenido Publicación de contenido Creación de contenido Ejecución Administración de fuentes de datos
Analista de negocio	pat	Publicación de contenido
Usuario avanzado	suzy	Lectura de contenido Publicación de contenido Acceso a contenido por horario Creación de contenido Ejecución
Ejecutor de reportes	tiffany	Publicación de contenido Acceso a contenido por horario

Pentaho BI Server 6.1 permite las siguientes funcionalidades de administración: adicionar usuarios, cambiar contraseñas de usuarios, eliminar usuarios, asignar usuarios a los roles, adicionar roles, asignar permisos a los roles, eliminar roles y asignar roles a los usuarios.

1.10 Mecanismos de autenticación

Según se plantea en (GSI - Facultad de Ingeniería, 2016) existen dos razones para autenticar a los usuarios de un sistema:

- La identidad del usuario es un parámetro para la decisión de control de acceso.

[10] Es un *framework* que se enfoca en proveer autenticación y autorización a las aplicaciones Java.

[11] Por sus siglas en inglés de NT LAN Manager.

- La identidad del usuario es registrada cuando se hace el *login* de eventos relevantes para la seguridad en la auditoría.

Cuando un usuario se conecta a un sistema de computadoras el mismo debe proveer usuario (paso de identificación) y contraseña (paso de autenticación). Siendo la autenticación el proceso de verificar una supuesta identidad y la verificación de identidades, es una o más de las siguientes opciones:

Algo que se sabe (ejemplo: password)

Algo que se tiene (ejemplo: badge, token, smart card)

Algo que se es (ejemplo: Huella digitales, ADN, iris)

Donde se está (ejemplo: Usando una terminal particular)

El proceso de autenticación consiste de varios pasos:

- Obtener la información de autenticación de una entidad.
- Analizar los datos.
- Determinar si la información de autenticación está efectivamente asociada a la entidad.

Existen un grupo de mecanismos de autenticación tales como: Passwords, Desafío-Respuesta, Mecanismos alternativos, Métodos múltiples, entre otros (GSI - Facultad de Ingeniería, 2016).

1.10.1 Servicio Central de Autenticación

Numerosas solicitudes de contraseña y el uso de diferentes credenciales para cada sistema han creado la necesidad de que las instituciones y organizaciones adopten un único proceso de autenticación de inicio de sesión web segura. El servicio central de autenticación de código abierto (CAS) proporciona una manera segura para que los usuarios accedan a múltiples servicios de una empresa. Fue creado originalmente por la universidad de Yale para proporcionar una forma de confianza en la autenticación de usuarios de una aplicación; convertido en un proyecto JASIG (actualmente Apereo[12]) en diciembre de 2004. Ofrece documentación a la comunidad y soporte de implementación, e incluye una extensa comunidad de adaptadores. El CAS está implementado como varios *Servlets* de Java, funciona a través del servidor HTTPS[13] y se accede a través de tres direcciones URL: la URL de *login*, la URL de validación y la URL de *logout*. Además, utiliza los *tickets*[14] como medio de autenticación, los cuales no pueden ser fácilmente falsificados ya que únicamente el servidor que los genera puede reconocerlos como válidos (Unicon, 2016).

En el escenario de CAS como se ilustra en la Figura 11, el navegador realiza una petición de servicio a una determinada aplicación (paso 1), esta aplicación busca el *ticket* del servidor CAS en la petición. Si el *ticket* no se encuentra es porque el usuario no se ha autenticado o su sesión ha caducado, en estos casos se redirige la petición hacia el sistema de autenticación de CAS (paso 2). Una vez realizado el proceso de autenticación correctamente, éste redirige al usuario nuevamente hacia la aplicación original pero incorporando el *ticket* correspondiente en la petición

[12] Fusión de las fundaciones JASIG y Sakai a finales de diciembre 2012.

[13] Hyper Text Transfer Protocol Secure en español: Protocolo seguro de transferencia de hipertexto.

[14] Número de caracteres único e irrepetible, generado por el servidor CAS.

(paso 3). Como la aplicación a la que se quiere acceder encuentra el *ticket* en la petición, pregunta a CAS (paso 4) si efectivamente el *ticket* es válido. Si la respuesta del CAS es afirmativa se permitirá al usuario acceder a esta aplicación. Si en el futuro, este usuario (y dentro de la validez de este *ticket*) intenta acceder a otra aplicación del sistema, esta solicitará al CAS que verifique el *ticket* del que el usuario ya dispone y le permitirá el acceso si la verificación es correcta. Una característica importante de esta arquitectura es que la única aplicación que conoce las credenciales de los usuarios es CAS, ya que el resto de aplicaciones sólo utiliza un *ticket* que tiene una caducidad temporal (Unicon, 2016).

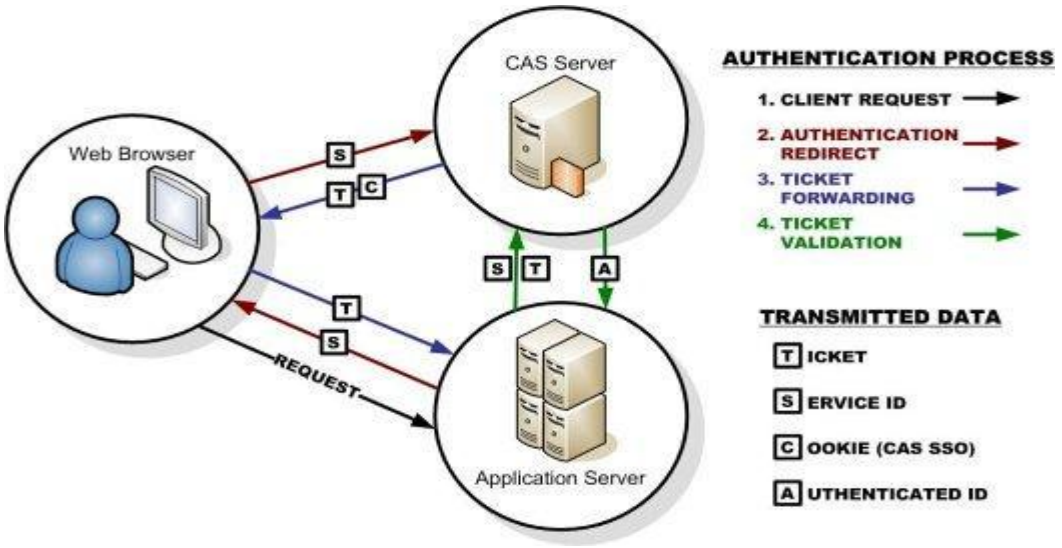


Figura 11: Sistema de autenticación CAS. (CAS overview, 2016).

Debido a las ventajas que trae consigo la implementación de un CAS para la adopción de un único proceso de autenticación, se analizó su posible integración como mecanismo de autenticación para realizar una propuesta de implementación a la herramienta Pentaho BI Server 6.1. Esta herramienta inicialmente no incluye el CAS pero su comunidad de desarrollo ofrece el procedimiento para incluirlo, lo cual debe ser realizado manualmente editando varios ficheros de configuración e incluyendo componentes necesarios.

1.10.2 Protocolo Ligero de Acceso a Directorios

El Protocolo Ligero de Acceso a Directorios (LDAP) es un conjunto de protocolos abiertos usados para acceder a la información guardada centralmente a través de la red. Está basado en el estándar X.500 para compartir directorios, pero es menos complejo e intensivo en el uso de recursos. LDAP organiza la información en un modo jerárquico usando directorios. Estos directorios pueden almacenar una gran variedad de información, permitiendo que cualquier usuario pueda acceder a su cuenta desde cualquier máquina en la red acreditada con LDAP.

Este protocolo representa un sistema cliente-servidor, donde el servidor puede usar una variedad de bases de datos para guardar un directorio, cada uno optimizado para operaciones de lectura rápidas y de gran volumen. Cuando una aplicación cliente LDAP se conecta a un servidor LDAP puede, o bien consultar un directorio, o intentar modificarlo. En el evento de una consulta, el servidor puede responderla localmente o puede dirigir la consulta a un servidor LDAP que tenga la respuesta. Si la aplicación cliente intenta modificar información contenida en un directorio LDAP, el servidor verifica que el usuario tiene permiso para efectuar el cambio y después añade o actualiza la información.

La mayor ventaja de LDAP es que se puede consolidar información para toda una organización dentro de un repositorio central. Puesto que LDAP soporta la capa de conexión segura (SSL) y la seguridad de la capa de transporte (TLS), los datos confidenciales pueden ser protegidos. También soporta un número de bases de datos *back-end* en las que se guardan directorios. Esto permite que los administradores tengan la flexibilidad de desplegar la base de datos más indicada para el tipo de información que el servidor tiene que controlar. Debido que LDAP posee una interfaz de programación de aplicaciones (API) bien definida, el número de aplicaciones acreditadas para LDAP son numerosas y están aumentando en cantidad y calidad (Understanding LDAP Design and Implementation, 2004).

Teniendo en cuenta las características y ventajas de LDAP tratadas anteriormente, se analizó su posible integración como mecanismo de autenticación para realizar una propuesta de implementación a la herramienta Pentaho BI Server 6.1. Esta herramienta inicialmente no incluye el LDAP pero su comunidad de desarrollo ofrece el procedimiento para incluirlo, lo cual debe ser realizado manualmente editando varios ficheros de configuración e incluyendo componentes necesarios.

1.11 Arquitectura Orientada a Servicios

La Arquitectura Orientada a Servicios (SOA) es un paradigma o estilo de arquitectura que se basa en la creación de un conjunto de servicios de diferente granularidad entre los procesos de negocio y las aplicaciones, con los siguientes objetivos (The server labs. The IT architects, 2016):

- Modelar la lógica de negocio como servicios para poder expresar la capa de negocio mediante la facilidad que ofrece la orquestación de los mismos.
- Crear una capa de servicios que ofrezca la funcionalidad de la capa de aplicación abstrayendo de la tecnología que la soporta.
- Minimizar las dependencias entre la capa de negocio y la de aplicación para desacoplar el negocio de la tecnología, y de este modo permitir los cambios en cualquiera de ellas. El objetivo es favorecer la agilidad para el negocio.
- Reutilizar los servicios de negocio creados en la organización, por medio de su publicación en el Bus de Servicios Corporativos (ESB-Enterprise Service Bus).

SOA posee varios niveles de servicios: servicios de aplicación, servicios de negocio y servicios de orquestación (The server labs. The IT architects, 2016), (Véase, Figura 12). Los servicios creados modelan la empresa según múltiples niveles de abstracción.

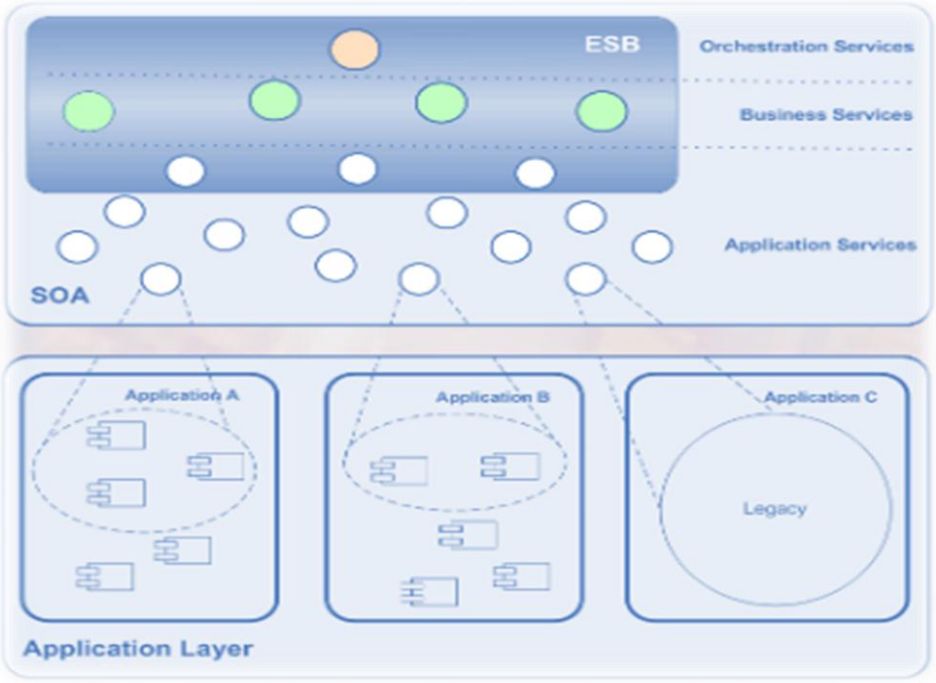


Figura 12: Niveles de servicios de SOA. (The server labs. The IT architects, 2016).

1.12 Gestión del conocimiento

¿Qué se entiende por Gestión del conocimiento? Varios son los autores que han establecido diversas definiciones de este concepto (Modelo de gestión del conocimiento de la investigación para Colombia y Cuba, 2012):

- Capacidad de la empresa para crear conocimiento nuevo, diseminarlo en la organización e incorporarlo en productos, servicios y sistemas. (Nonaka, I. y Takeuchi, año 1995).
- Proceso de creación de conocimiento, validación, presentación, distribución y aplicación de conocimiento (Bhatt, D. G, año 2001).
- Conjunto de actividades realizadas con el fin de buscar, utilizar, procesar, compartir y desarrollar conocimiento en una organización y en los individuos que en ella trabajan, encaminándolas a la mejor consecución de sus objetivos y a aprender (Torres, año 2009).
- Es cómo convertir el conocimiento en valor agregado para los clientes, la organización y la sociedad en su conjunto. (Franch, Antunez, Herrera, año 2012).
- Arreglo estructurado del conjunto de políticas, recursos, procesos, resultados y productos, que tienden a lograr un fin establecido, lo que le confiere un ordenamiento sistémico (Modelo de gestión del conocimiento de la investigación para Colombia y Cuba, 2012).

La Gestión del conocimiento está formada por un conjunto de actividades básicas, siendo las más generales las destinadas a identificar, crear, almacenar, compartir y utilizar el conocimiento (Véase, Figura 13) (Importancia de la utilización de un Data Warehouse (DW) en las empresas, 2006). Esta tecnología proporciona el almacenamiento de la información de forma tal que permite el descubrimiento de conocimiento partiendo de la misma.

Figura 13: Definición de Gestión del conocimiento. Elaboración propia.

El conocimiento se origina y aplica en la mente de las personas. En las organizaciones, el conocimiento reside en documentos, bases de datos y también en los procesos, prácticas y normas corporativas. La Figura 14 muestra un Modelo General del Conocimiento, cuyo flujo se representa en cuatro actividades básicas: creación, retención, transferencia y utilización del conocimiento (Metodología para la Extracción del Conocimiento Empresarial a partir de los Datos, 2006).

Figura 14: Modelo General del Conocimiento (Newman, 2000).

- **Creación del Conocimiento:** comprende las actividades asociadas con la entrada de nuevo conocimiento al sistema, e incluye el desarrollo, descubrimiento y captura del conocimiento.
- **Retención del Conocimiento:** incluye todas las actividades que preservan el conocimiento y que permiten mantenerlo en el sistema.
- **Transferencia del conocimiento:** se refiere a las actividades asociadas con el flujo del conocimiento desde una parte a otra. Incluye la comunicación, traslación, conversión, filtrado y suministro.
- **Utilización del Conocimiento:** abarca las actividades relacionadas con la aplicación del conocimiento a los procesos de negocio.

El conocimiento se puede clasificar en explícito y tácito. El conocimiento explícito es el que está almacenado en algún medio, se encuentran establecidos los procedimientos por los cuales se puede trasmitir a otras personas. Por el contrario, el conocimiento tácito está relacionado absolutamente a las personas, es aquel que poseen las organizaciones, pero que no se plasma ni se registra. (Véase, Figura 15).

Figura 15: Clasificación del conocimiento. Elaboración propia.

Gestionar de forma adecuada el conocimiento en las empresas constituye un activo importante, pues les permite obtener ventaja competitiva en el mercado. Entre los objetivos de la Gestión del conocimiento se encuentran:

- Formular una estrategia de alcance organizacional para el desarrollo, adquisición y aplicación del conocimiento.
- Implantar estrategias orientadas al conocimiento.
- Promover la mejora continua de los procesos de negocio, enfatizando la generación y utilización del conocimiento.
- Monitorizar y evaluar los logros obtenidos mediante la aplicación del conocimiento.
- Reducir los tiempos de ciclo en el desarrollo de nuevos productos y mejorar los existentes.
- Reducir los costos por repetición de errores.

En la investigación se procura utilizar la Gestión del conocimiento para gestionar tanto conocimiento tácito como explícito relacionado con los AD, fundamentales para el establecimiento de los componentes de la estrategia de seguridad que se propone.

1.13 Características de la estrategia científica

Los autores en (de Armas Ramírez, y otros, 2001) ofrecen puntos de vista acerca de la definición y diseño de la estructura de diferentes tipos de resultados científicos de la investigación y reflexionan acerca de los procedimientos lógicos y metodológicos que subyacen en la construcción de las propuestas. Particularmente, a continuación se presentan las principales características de la estrategia:

- Se **diseñan** para resolver problemas de la práctica y vencer dificultades con optimización de tiempo y recursos.
- **Permiten** proyectar un cambio cualitativo en el sistema al eliminar las contradicciones entre el estado actual y el deseado.
- **Implican** un proceso de planificación en el que se produce el establecimiento de secuencias de acciones orientadas hacia el fin a alcanzar, lo cual no significa un único curso de las mismas.
- **Interrelacionan** dialécticamente en un plan global los objetivos o fines que se persiguen y la metodología para alcanzarlos.

Los principales elementos expuestos anteriormente se pueden esquematizar de la siguiente forma:

Figura 16: Principales elementos de la estrategia. (de Armas Ramírez, y otros, 2001).

Toda estrategia transita por una acción de obtención de información (puede tener carácter diagnóstico), una acción de utilización de información y una acción de evaluación de esa información, además como su nombre lo indica, debe tener un margen para ir redirigiendo las acciones. La estrategia establece la dirección inteligente, y desde una perspectiva amplia y global, de las acciones encaminadas a resolver los problemas detectados en un determinado segmento de la actividad humana. Se entienden como problemas las contradicciones o discrepancias entre el estado actual y el deseado, entre lo que es y debería ser, de acuerdo con determinadas expectativas. Su diseño implica la articulación dialéctica entre los objetivos (metas perseguidas) y la metodología (vías instrumentadas para alcanzarlas) (de Armas Ramírez, y otros, 2001).

1.14 Conclusiones parciales

Partiendo del análisis realizado hasta el momento, se puede concluir que:

El estudio del marco teórico de la investigación relacionado con el proceso de implementación de mecanismos de seguridad en diferentes arquitecturas para la construcción de almacenes de datos, permitió conocer que: en la literatura existen propuestas para la inclusión de la seguridad en elementos específicos presentes en la arquitectura general de los almacenes de datos, sin embargo, no se consideran aspectos de seguridad que deben incorporarse en todas las etapas del ciclo de desarrollo.

El estudio de la situación actual de las estrategias de seguridad en las soluciones de almacenes de datos del centro DATEC de la UCI permitió concluir que:

- La metodología de desarrollo utilizada establece las actividades, fases que deben desarrollarse para implementar los requisitos de seguridad y los artefactos como resultado de su implementación, sin embargo, esta metodología no aborda los procedimientos necesarios que deben guiar dicha especificación e implementación.
- La mayoría de los proyectos que se desarrollan se limitan a garantizar la seguridad en el diseño físico del despliegue de la solución, sin tener en cuenta la seguridad necesaria en los subsistemas presente en la arquitectura utilizada.
- La arquitectura definida no organiza la información almacenada en los AD, en relación con el valor de los datos y la gravedad de las amenazas a las que puedan estar sometidos. Esto imposibilita que se puedan definir diferentes mecanismos de seguridad, teniendo en cuenta dicha relación.

- La configuración básica de la herramienta Pentaho BI Server 6.1 ofrece funcionalidades limitadas respecto a la seguridad de la información, debido que: los mecanismos de autenticación se especifican de forma manual editando directamente los ficheros de configuración, las contraseñas de los usuarios no se validan eficazmente y no existe la posibilidad de restringir el acceso mediante direcciones IP.
- Inconvenientemente, no se encontró una estrategia para elevar la seguridad en todo el ciclo de desarrollo de las soluciones de AD del centro DATEC de la UCI.

Teniendo en cuenta los elementos anteriores, es necesaria una solución que permita elevar la calidad de este tipo de soluciones, a través de la seguridad de los elementos presentes en su desarrollo. En este sentido la autora del presente trabajo se inclina por desarrollar una estrategia de seguridad que dirija las acciones de los especialistas de AD en el proceso de implementación de mecanismos de seguridad; utilizando como herramienta la Gestión del conocimiento.

CAPÍTULO II. DESCRIPCIÓN DE LA ESTRATEGIA DE SEGURIDAD PROPUESTA PARA SOLUCIONES DE ALMACENES DE DATOS

En el presente capítulo se desarrolla una estrategia de seguridad con el objetivo de elevar la protección de los datos en cada subsistema de las soluciones de AD del centro DATEC de la UCI. Como parte de la estrategia se definen las acciones a seguir en el proceso de implementación de mecanismos de seguridad, se abordan los principios de seguridad necesarios, se exponen ejemplos de mecanismos de seguridad para cada subsistema y se propone una arquitectura de seguridad.

2.1 Propuesta de solución

El análisis de los principales elementos para definir el tipo de resultado científico de la presente investigación, permitió establecer la propuesta de una estrategia de seguridad. En su diseño se aplicaron métodos de investigación científica, entre otras razones por sus objetivos, que son:

- Representar una guía que encamine las acciones coordinadas de los especialistas de AD para gestionar la seguridad de este tipo de soluciones.
- Establecer un ciclo de acciones de mejora continua para elevar la calidad del AD, a través de su seguridad.
- Seleccionar los mecanismos de seguridad más apropiados para elevar la seguridad de los AD.
- Abordar los procedimientos necesarios para la implementación de mecanismos de seguridad en los AD.
- Obtener uniformidad en el proceso de implementación de mecanismos de seguridad en soluciones de AD.
- Ofrecer métodos y ejemplos de soluciones para gestionar la seguridad de AD.
- Obtener mejores indicadores en las soluciones de AD que se desarrollen.

Para la definición de la estrategia de seguridad para soluciones de almacenes de datos en el presente trabajo se utilizó la metodología de Gestión del conocimiento definida en (Modelo de gestión del conocimiento de la investigación para Colombia y Cuba, 2012), cuyas fases se muestran en la Figura 17.

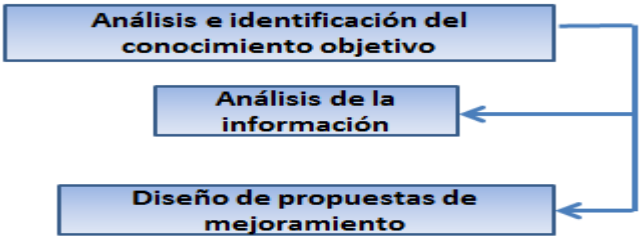


Figura 17: Fases de la metodología de Gestión del Conocimiento definida en (Modelo de gestión del conocimiento de la investigación para Colombia y Cuba, 2012).

Análisis e identificación del conocimiento objetivo: en esta etapa se realizó el levantamiento de los elementos de seguridad críticos que debía contener la estrategia de seguridad, así como de las vulnerabilidades y riesgos relacionados con cada subsistema, incluyendo las herramientas de desarrollo utilizadas. Esta información se recogió a través de encuestas y entrevistas efectuadas a especialistas en el desarrollo de almacenes de datos del centro DATEC de la UCI. La muestra

seleccionada fue de 30 especialistas. Para el análisis se utilizaron además los métodos de investigación acción, el análisis de campos de fuerzas y el diagrama causa-efecto.

Análisis de la información: se realizó la tabulación de las encuestas y el análisis de las entrevistas a los diferentes actores involucrados. Esto permitió evaluar las vulnerabilidades, los riesgos relacionados con cada subsistema, y la criticidad de los elementos de seguridad identificados en la fase anterior.

Diseño de propuestas de mejoramiento: esta fase culminó con la elaboración de la estrategia de seguridad para las soluciones de almacenes de datos del centro DATEC de la UCI.

2.2 Estrategia de seguridad para almacenes de datos

La estrategia de seguridad que se propone incluye las acciones a seguir para la aplicación de mecanismos de seguridad a cada uno de los subsistemas que se desarrollan, aborda principios de seguridad, mecanismos de seguridad y una propuesta de arquitectura de seguridad. Los elementos que integran la estrategia de seguridad que se propone se representan en la Figura 18.

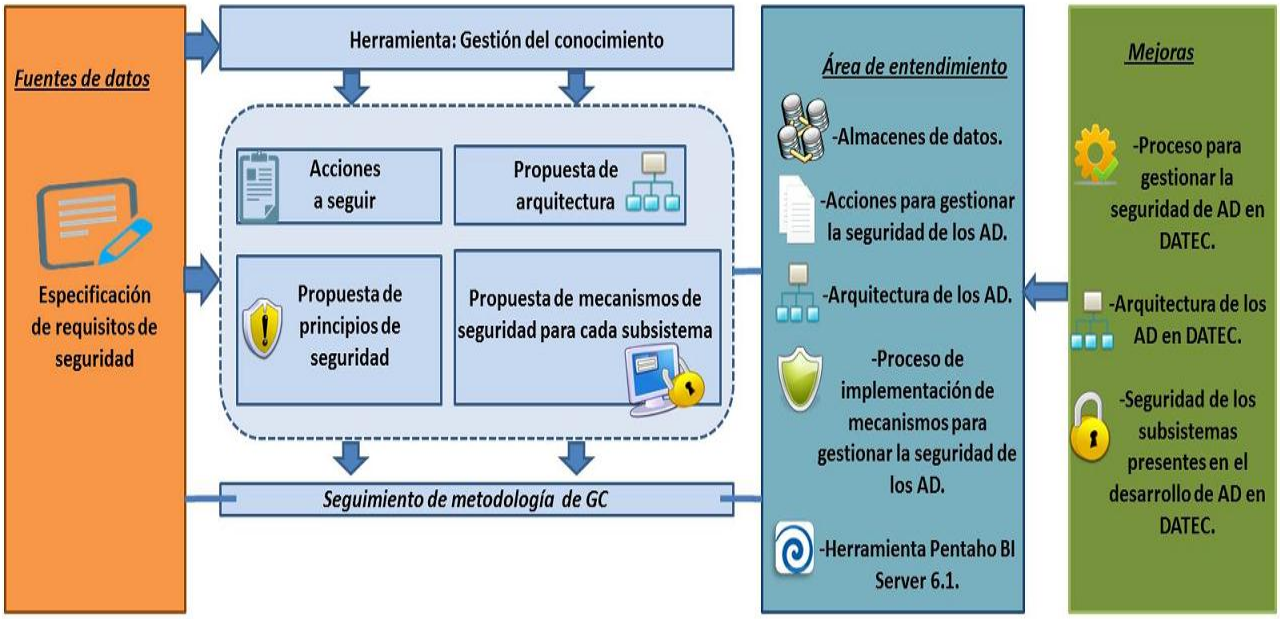


Figura 18: Componentes de la estrategia de seguridad que se propone, basados en (Arquitecturas de Nueva Generación , 2016) . Elaboración propia.

Se utilizó como herramienta la Gestión del conocimiento, sustentada en el seguimiento de la metodología de Gestión del conocimiento utilizada en (Modelo de gestión del conocimiento de la investigación para Colombia y Cuba, 2012); para lo cual se gestionó tanto conocimiento tácito como explícito.

La seguridad de la información es un requisito fundamental que debe ser pensado cuidadosamente y debe estar presente en todas las etapas del ciclo de desarrollo de un almacén de datos. La selección de los mecanismos de seguridad debe realizarse de acuerdo a las necesidades específicas de la institución y después de haber realizado un análisis de las acciones necesarias, las cuales deben basarse fundamentalmente en los factores que hacen a estas soluciones particularmente susceptible a ataques, como se describió en el capítulo 1; u otros factores necesarios que sean identificados por la institución.

Para la aplicación de la estrategia que se propone se necesita como entrada el artefacto especificación de requisitos de seguridad para cada subsistema del almacén de datos. Este indica las características de seguridad que el almacén de datos debe poseer. Es esencial que una organización identifique sus requisitos de seguridad. La Tabla 10 del Anexo 1 muestra formato y ejemplo para la especificación de requisitos que se necesita. Existen tres fuentes principales para la especificación de los requisitos de seguridad (27001, 2005):

- La evaluación de los riesgos para la organización, teniendo en cuenta las estrategias de negocios y objetivos de la organización. A través de una evaluación de riesgos se identifican las amenazas a los activos, se evalúa y se estima el impacto potencial de la probabilidad de ocurrencia de las vulnerabilidades.
- Se tienen que cumplir los requisitos legales, estatutarios, reglamentarios y contractuales de la organización, de los socios comerciales, contratistas y proveedores de servicios, y su entorno socio-cultural.
- Tienen que desarrollarse los principios, objetivos y requisitos de negocio para el tratamiento, procesamiento, almacenamiento y comunicación de la información; para dar respuesta a las operaciones de una organización.

Se realiza la evaluación de los riesgos según el monitoreo guiado por listas de chequeo, evaluando el riesgo de cada parámetro existente en las mismas. Existen herramientas que automatizan esta evaluación de riesgos entre las que se encuentran las que incorporan componentes del protocolo de automatización de contenido de seguridad SCAP; el cual constituye una iniciativa del Instituto Nacional de Estándares y Tecnologías de los EEUU (NIST por sus siglas en inglés) para permitir la interoperabilidad de diferentes aplicaciones y la adopción de un lenguaje de seguridad. Las organizaciones deberían emplear listas de chequeo de configuraciones de seguridad expresadas en SCAP para mejorar y monitorear la seguridad de los sistemas (Montesino Perurena, 2013). De igual manera existen metodologías para la gestión de los riesgos, entre las que se encuentran: MGERIT, OCTAVE, CRAMM, COBIT y otras.

La estrategia de seguridad que se propone considera un grupo de controles del estándar ISO/IEC 27001, relacionados con dominios tales como: políticas de seguridad, organización de la seguridad de la información, administración de activos, clasificación de la información, seguridad física y ambiental, respaldo de la información, gestión de medios, intercambio de la información, control de acceso y adquisición, desarrollo y mantenimiento de sistemas de información.

2.2.1 Acciones de la estrategia de seguridad

Las acciones a seguir que propone la presente investigación son las siguientes:

Acción 1: Identificación de los datos.

Se debe realizar un inventario completo de todos los datos, tales como: la fuente de datos y su medio de almacenamiento, el volumen de la información, los metadatos que se utilizarán en los procesos de integración, los datos que se almacenarán en el área de almacenamiento intermedio conocida como "*staging area*", entre otros. La Tabla 11 del Anexo 1 muestra formato y ejemplo para registrar el resultado de esta acción.

Acción 2: Clasificación de los datos.

La clasificación de todos los datos es necesaria para satisfacer los requisitos de seguridad relacionados con la confidencialidad, integridad y disponibilidad de la información. La realización de esta acción requiere de la participación de todas las partes interesadas y dicha clasificación se realiza de manera conciliada. Los datos se clasifican generalmente sobre la base de la criticidad o sensibilidad a la divulgación, modificación, y destrucción de la información en el entorno donde está enmarcada la solución. Se establece un esquema de clasificación específico para cada empresa, basado en que tan crítica y sensible es la información. Se clasifican los datos identificados según su nivel de sensibilidad en: públicos los menos confidenciales, confidenciales los datos moderadamente sensibles y secretos los más sensibles. Por ejemplo:

- **Datos públicos** pudieran ser: los datos relativos al estado civil de las personas, a su profesión u oficio y a su calidad de comerciante o de servidor público. Por su naturaleza, los datos públicos pueden estar contenidos en registros públicos, documentos públicos, gacetas, boletines oficiales y sentencias judiciales debidamente ejecutadas que no estén sometidas a reserva, entre otros.
- **Datos confidenciales** pudieran ser: datos personales, entre ellos información de identificación personal, como los números de documento nacional de identidad o del seguro social, números de pasaporte, números de tarjeta de crédito, entre otros datos personales. Registros financieros, incluidos los números de cuentas financieras, como números de cuentas corrientes o de inversión. Material de negocios, como documentos o datos exclusivos o propiedad intelectual específica, entre otros.
- **Datos secretos**: se entiende por datos secretos aquellos que afectan la intimidad de las personas o cuyo uso indebido puede generar su discriminación, tales como aquellos que revelen el origen racial o étnico, la orientación política, las convicciones religiosas o filosóficas, la pertenencia a sindicatos, organizaciones sociales, de derechos humanos o que promueva intereses de cualquier partido político, así como los datos relativos a la salud, a la vida sexual, y los datos biométricos, entre otros. La Tabla 12 del Anexo 1 muestra formato y ejemplo para registrar el resultado de esta acción.

Independientemente que estas clasificaciones se realizan de manera conciliada con el cliente dependiendo de las características de su información, existen normas internacionales específicas para la clasificación de los datos que pueden utilizarse como guía. Por ejemplo las normas establecidas en (Presidencia de la República, 2016).

Acción 3: Identificación de las vulnerabilidades en la seguridad de los datos.

Se identifican y documentan las vulnerabilidades relacionadas con el medio de almacenamiento de la fuente de datos, la comunicación y el acceso, las vulnerabilidades de las herramientas técnicas que se utilizarán en el desarrollo de la solución, la arquitectura general del AD, el medio ambiente teniendo en cuenta la administración de los grupos de usuarios y las contraseñas, la distribución física de los servidores, los sistemas operativos a utilizar, la identificación de las vulnerabilidades de los ficheros generados en el desarrollo de cada subsistema. La Tabla 13 del Anexo 1 muestra formato y ejemplo para registrar el resultado de esta acción.

Existen herramientas que pueden utilizarse para identificar vulnerabilidades, por ejemplo CVE Details (www.cvedetails.com, 2016). Esta herramienta permite consumir datos de las

vulnerabilidades y exposiciones comunes registradas por la base de datos nacional de vulnerabilidades del NIST del gobierno de los EE UU (González Brito, 2016).

Acción 4: Identificación de medidas de protección de datos.

Se identifican medidas de seguridad para cada una de las vulnerabilidades identificadas, teniendo en cuenta la sensibilidad de la información. Para la realización de esta acción se deberán considerar los principios de seguridad expuestos en el próximo epígrafe. De igual forma se considera necesario realizar un estudio de los mecanismos existentes según las necesidades de cada solución; principalmente de los expuestos en la presente investigación. La Tabla 14 del Anexo 1 muestra formato y ejemplo para registrar el resultado de esta acción.

Acción 5: Implementación de medidas de protección de datos.

Se implementan las medidas de protección de datos identificadas en la acción anterior.

Acción 6: Evaluación de la efectividad de las medidas de seguridad.

Se propone diseñar e implementar casos de pruebas que traten de ejecutar las vulnerabilidades identificadas, y posteriormente realizar un análisis de los resultados obtenidos. La Tabla 15 del Anexo 1 muestra formato y ejemplo para registrar el resultado de esta acción.

Acción 7: Tomar acciones para mejorar continuamente el desempeño de la gestión de seguridad.

Se propone realizar: seguimientos a las acciones para gestionar la seguridad de las soluciones de AD, evaluación del cumplimiento, investigaciones de incidencias, análisis de las no conformidades detectadas, acciones correctivas y preventivas, auditorías internas, control de riesgos, entre otras.

El diagrama representado en la Figura 19 destaca las acciones que guiarán a los especialistas de AD del centro DATEC en el proceso de implementación de mecanismos de seguridad. Además, se especifican los artefactos de entrada y salida de cada acción.

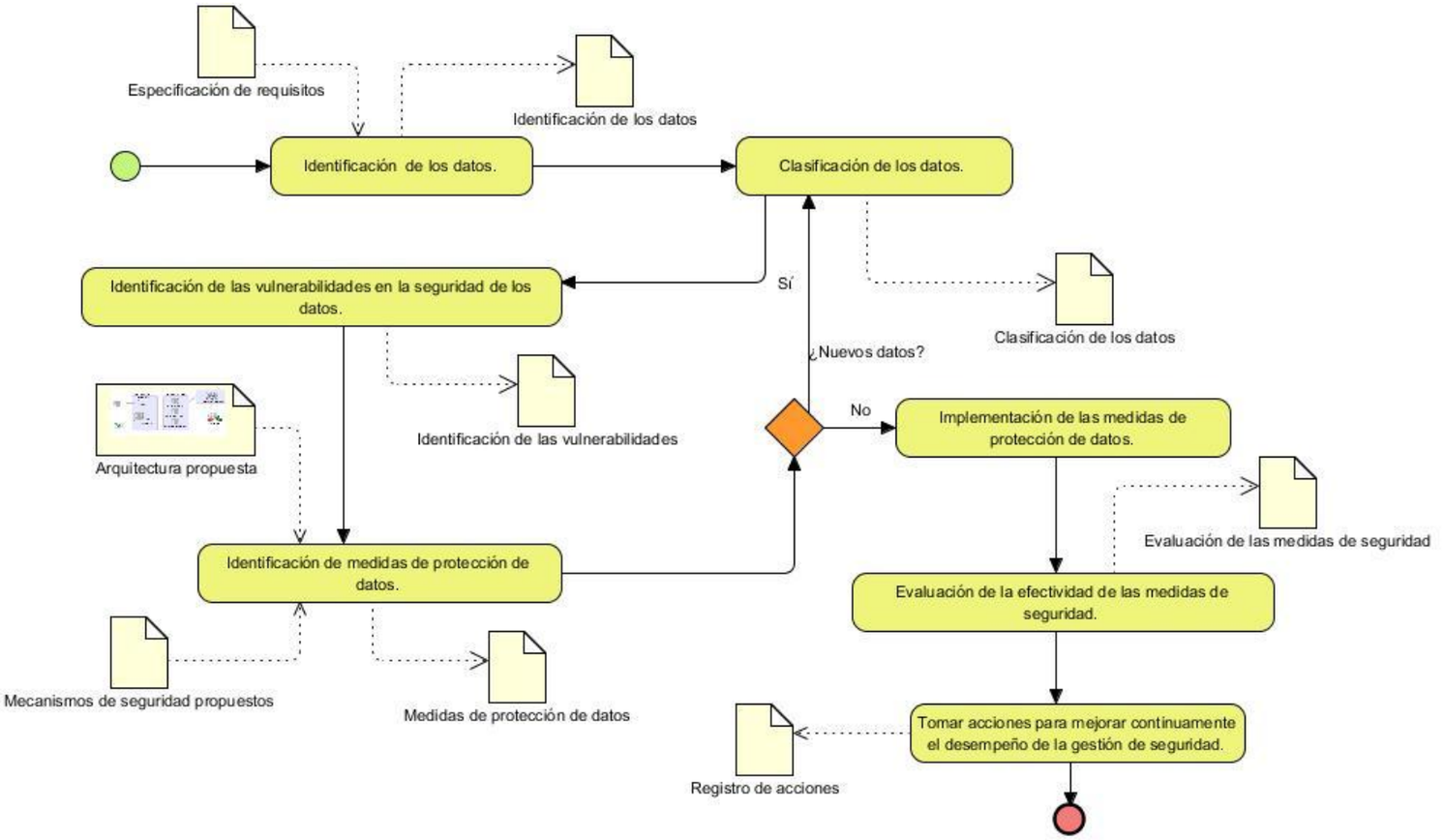


Figura 19: Diagrama de actividades de la estrategia de seguridad que se propone. Elaboración propia.

Aunque todos los especialistas estarán involucrados en la realización de estas acciones, se deberá designar un responsable dentro de cada grupo de desarrollo (Almacenes, Integración de datos e Inteligencia de Negocio) del proyecto para dirigir las mismas. La Figura 20 muestra cómo se organizan las acciones en cuatro etapas cíclicas: Planificar, Hacer, Verificar y Actuar siguiendo un círculo PDCA.

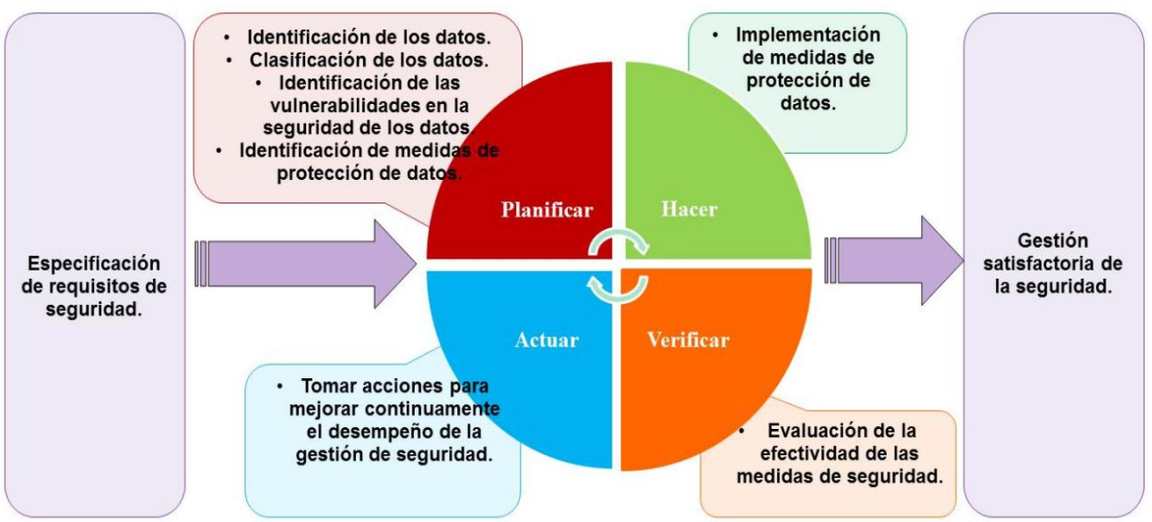


Figura 20: Organización de las acciones siguiendo un círculo PDCA. Elaboración propia.

El círculo PDCA tiene la siguiente interpretación: cuando se busca obtener algo, lo primero que hay que hacer es planificar cómo conseguirlo (planificar), después se procede a realizar las acciones planificadas (hacer), a continuación se comprueba lo que se ha hecho (verificar) y finalmente se implementan los cambios pertinentes para no volver a incurrir en los mismos errores (actuar). Nuevamente se comienza con la ejecución del ciclo planificando introduciendo las mejoras provenientes de la experiencia anterior.

2.2.2 Principios de la estrategia de seguridad

Después de realizar un análisis acerca de algunos principios de seguridad para soluciones de almacenes de datos y para la seguridad de la información de manera general, la estrategia de seguridad que se propone considera los siguientes principios:

- Principio de menor privilegio.
- Seguridad no equivale a oscuridad.
- Principio del eslabón más débil.
- Defensa en profundidad.
- Punto de control centralizado.
- Seguridad en caso de fallo.
- Participación universal.
- Principio de simplicidad.
- Arquitectura distribuida.
- Copias diarias hacia medios seguros.
- Situar estratégicamente *gateway* de filtrado de paquetes.
- Cuello de botella de la autenticación y el acceso.
- Integridad y validación de los datos.
- Enmascaramiento de datos y conservación de la privacidad.
- Políticas de accesos y restricciones de datos.
- Clasificación de los datos.

En el capítulo 1 se realizó una explicación de cada uno de estos principios de seguridad y su importancia, los mismos deberán ser considerados principalmente en la acción 4 (Identificación de medidas de protección de datos) de la estrategia de seguridad que se propone, pues son necesarios para diseñar cualquier medida de seguridad.

2.2.3 Propuesta de arquitectura de seguridad

Considerando que los principios y mecanismos son descritos en la bibliografía, si bien de manera general, o enfocados a una arquitectura específica, o a modelos específicos, como se describió en el capítulo 1; a continuación (Véase, Figura 21) se presenta una propuesta de arquitectura de seguridad para las soluciones de almacenes de datos del centro DATEC de la UCI, que responde a sus características específicas. Esta arquitectura constituye una adaptación siguiendo varios de los principios expuestos en el presente trabajo y dará soporte a los mecanismos de seguridad que se definirán para cada uno de sus subsistemas que se desarrollan.

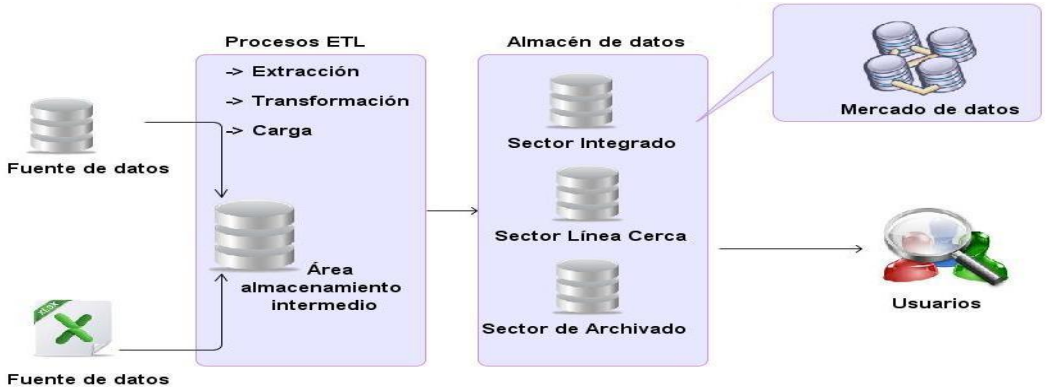


Figura 21: Arquitectura de seguridad que se propone. Elaboración propia.

El principio básico de la arquitectura que se propone es el principio de arquitectura distribuida destacado por Kimball. La implementación sobre una arquitectura con sectores independientes reduce en gran medida la vulnerabilidad de los AD ante ataques y fallos de un solo punto; pues permite establecer mecanismos de seguridad en dependencia de las características específicas de cada sector. Se siguió la concepción de una arquitectura *"bottom-up"* o ascendente del problema, que incorpora elementos propuestos por Inmon en (Inmon, y otros, 2008), adaptándolos a las necesidades específicas de las soluciones de AD del centro DATEC. A continuación se explican los elementos que conforman la arquitectura:

Los subsistemas Fuente de datos y Procesos ETL mantienen los mismos criterios que se presentaron en el capítulo 1. Por otro lado, se propone que la construcción de los mercados de datos que conforman el AD se desarrolle sobre tres sectores o subsistemas: **Integrado**, **Línea Cerca** y **Archivado** similar a los propuestos por Inmon en (Inmon, y otros, 2008).

Sector Integrado

La función del sector **Integrado** es similar a la del sector **Interactivo** de Inmon. La diferencia fundamental es que los datos son operacionales, almacenados después de realizar las actividades de ETL tanto para datos estructurados o no estructurados. Este sector almacena los datos de mayor acceso y el volumen es poco porque solo se tiene la información de los procesos actuales de la empresa, por lo que el tiempo de respuesta a solicitudes realizadas a este sector es de pocos segundos.

Sector Línea Cerca

Al igual que el sector **Línea Cerca** de Inmon, este sector se implementará de manera opcional cuando exista mayor cantidad de datos y la probabilidad de acceso sea menor.

Sector Archivado

A diferencia del sector **Archivado** de Inmon que propone registrar la información histórica a partir de los 5 años, se propone almacenar en este sector la información de otros procesos dentro de la empresa exceptuando el actual. Convirtiéndose en este momento en datos históricos de la institución.

La Tabla 2 describe brevemente en qué consiste cada sector propuesto.

Tabla 2: Descripción de los sectores del subsistema Almacén de datos. Elaboración propia.

Sector Integrado	Sector Línea Cerca	Sector Archivado
Datos operacionales, almacenados después de realizar las actividades de ETL. Almacena los datos de mayor acceso y el volumen es poco.	Se implementará de manera opcional cuando exista un gran volumen de datos. Se trasladarán a este sector los datos de menor probabilidad de acceso.	Almacenará la información de otros procesos dentro de la empresa exceptuando el actual, convirtiéndose en datos históricos de la institución.

2.2.4 Mecanismos de seguridad a nivel de subsistema

Seguridad. Subsistema Fuente de datos

Las fuentes de datos para el desarrollo de un almacén de datos son de naturaleza muy diversa (ficheros de texto plano, sistemas operacionales como bases de datos jerárquica, relacional, entre otros). Para asegurar la fuente de datos, los mecanismos a utilizar dependen del tipo de fuente. Por ejemplo, si la fuente es una base de datos, deben establecerse políticas de administración de grupos de usuarios y contraseñas para el acceso. Si la fuente de datos constituyen ficheros de texto plano, se deberán proteger mediante mecanismos de seguridad desde el sistema operativo, como por ejemplo el módulo de seguridad SELinux (The SELinux Notebook - The Foundations, 2007), el uso de Listas de Control de Acceso (ACLs) en Linux, entre otros que ofrecen seguridad mediante el control de acceso.

Se recomienda realizar respaldo de las fuentes de datos hacia medios seguros, y de igual manera utilizar mecanismos de seguridad para su protección. Las herramientas a utilizar para la limpieza de los datos deberán implementar mecanismos de autenticación mediante usuario y contraseña, para proveer el acceso solo a los especialistas de ETL. Debido que, si se lleva a cabo una mala ejecución de las reglas de limpieza, esto puede ocasionar modificaciones y pérdidas no deseadas.

La Figura 22 resume los mecanismos que deben tenerse en cuenta de manera obligatoria para asegurar el subsistema Fuente de datos.

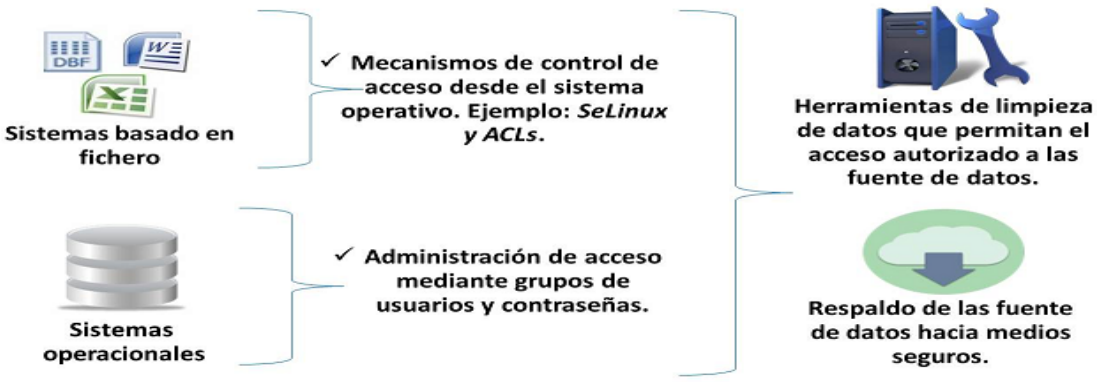


Figura 22: Ejemplos de mecanismos de seguridad para el subsistema Fuente de datos. Elaboración propia.

Seguridad. Subsistema Procesos ETL

En el subsistema de los procesos ETL se deben almacenar en forma cifrada los metadatos utilizados, principalmente aquellos metadatos que especifican parámetros de conexión, e

identificación de aspectos confidenciales de la empresa. Por otro lado, solo los especialistas de ETL deben tener acceso total a la información que se almacena en dicha área y en el área de "*staging area*", como plantea Kimball en (Kimball, y otros, 2004), (Véase, Figura 23).

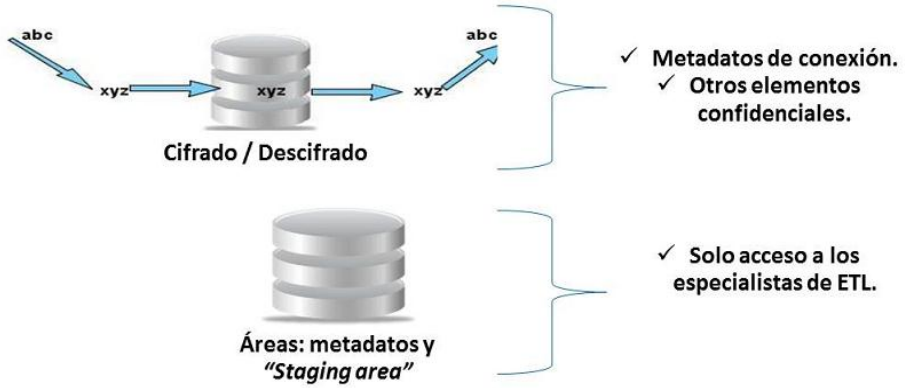


Figura 23: Ejemplos de mecanismos de seguridad para el subsistema Procesos ETL. Elaboración propia.

La comunicación entre la fuente de datos y el subsistema Procesos ETL se debe realizar a través del repositorio que provee la herramienta Pentaho Data Integration 6.1, utilizada por los desarrolladores ETL en el proceso de integración. Para acceder al repositorio se necesitan cuentas de usuario asociadas a perfiles que determinen sus permisos. Existen tres perfiles predefinidos: usuarios con permisos de solo lectura, usuario con todos los permisos para trabajar con los diferentes objetos y el usuario administrador, que tiene también todos los permisos; pero además puede incluir y eliminar nuevos usuarios y perfiles.

Se deberán utilizar políticas específicas según la fuente de datos, como por ejemplo: si la fuente de datos es un repositorio de ficheros de texto plano se podrá verificar la integridad de la información, si se envía al subsistema de procesos ETL el *hash* de cada fichero, y en la herramienta de integración se podrá comprobar con dicho *hash* si se ha modificado el contenido del mismo.

Seguridad. Subsistema Almacén de datos

En el subsistema Almacén de datos, los sectores: **Integrado**, **Línea Cerca** y **Archivado**, pueden ser accedidos directamente desde el servidor de aplicaciones, según las necesidades de información de los usuarios finales. Por consiguiente, se propone aislar el servidor de base de datos compuesto por estos tres subsistemas en una red de área local virtual o más conocida por sus siglas en inglés como VLAN; y ubicarlo detrás de un *gateway* de filtrado de paquetes, de manera que solo pueda recibir paquetes desde el exterior si proceden del servidor Procesos ETL, (Véase, Figura 24).

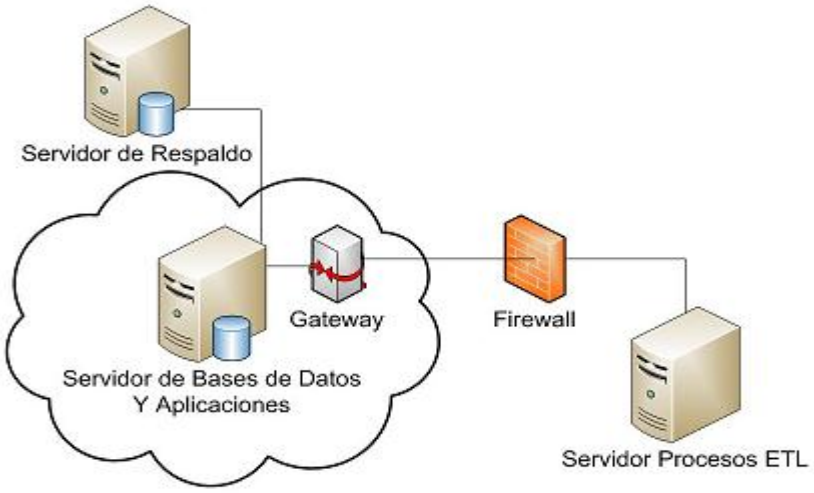


Figura 24: Despliegue para servidores de Bases de Datos y Procesos ETL. Elaboración propia.

Se deberá además:

Implementar un modelo de control de acceso basado en roles también conocido como RBAC (por sus siglas en inglés: Role Based Access Control), para proteger los objetos del AD. Este modelo es uno de los esquemas más representativos de control de acceso, el cual contiene tres procesos fundamentales: autenticación, autorización y auditoría (RBAC Extension Model for ERP Systems in Multi-domain Environments, 2012).

La Tabla 3 representa un ejemplo de roles y permisos para controlar el acceso a un almacén de datos.

Tabla 3: Ejemplo de roles y permisos para controlar el acceso al almacén de datos. Elaboración propia.

Roles	Permisos
Administrador de base de datos	Total acceso a la base de datos. Lleva a cabo la política de respaldo y recuperación.
Administrador de ETL	Permiso de lectura y escritura sobre esquemas de la base de datos que contengan información de negocio y todos los permisos sobre los esquemas metadatos y "*staging area*".
Administrador	Acceso total al área de análisis a nivel de aplicación de visualización.
Analista	Tiene permiso de solo lectura en el acceso a las áreas de análisis, que incluyen los libros de trabajo de los reportes de la aplicación.

También es necesario controlar la autenticación de clientes al sistema gestor de bases de datos PostgreSQL que se utiliza; mediante la manipulación de los parámetros del fichero pg_hba.conf. Se deberán establecer políticas de seguridad que pudieran ser diferentes en cada sector de la arquitectura propuesta, como por ejemplo: el cifrado de datos en el sector **Integrado**, por ser el sector con mayor probabilidad de acceso y contener menor volumen de los datos, así como realizar salvas periódicas de la información histórica almacenada en el sector **Archivado** y del

resto de los sectores (Véase, Figura 25). Además, podrán implementarse otros mecanismos de seguridad que se explican en la investigación.

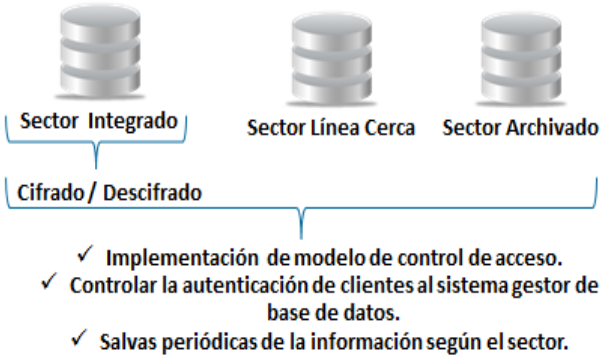


Figura 25: Ejemplos de mecanismos de seguridad para subsistemas de la arquitectura que se propone. Elaboración propia.

Seguridad. Subsistema Visualización de los datos

Los requisitos funcionales son capacidades o condiciones que un sistema determinado debe cumplir. Expresan una especificación detallada de las responsabilidades del sistema en cuestión y permiten determinar, de una manera clara, lo que debe realizar el sistema (Pressman, 2007). Dadas las limitaciones respecto a la seguridad en las opciones de la configuración básica de la herramienta Pentaho BI Server 6.1, se identifican en la investigación 15 requisitos funcionales los cuales se definen a continuación:

RF1 Autenticar usuario.

RF1.1 Validar la complejidad de las contraseñas.

RF1.2 Controlar el historial de las contraseñas.

RF2 Configurar de forma segura los gestores de bases de datos.

RF2.1 Configurar de forma segura el gestor de bases de datos PostgreSQL.

RF2.2 Configurar de forma segura el gestor de bases de datos MySQL.

RF2.3 Configurar de forma segura el gestor de bases de datos Oracle.

RF3 Integrar mecanismos de acceso con LDAP.

RF3.1 Validar la conexión a la base de datos.

RF4 Restringir conexiones de usuario.

RF4.1 Restringir conexión por una dirección IP específica.

RF4.2 Restringir conexión por un rango de direcciones IP determinado.

RF9 Configurar la conexión con el CAS.

RF10 Gestionar conexiones de bases de datos.

RF10.1 Listar bases de datos.

Implementación propuesta para la herramienta Pentaho BI Server 6.1

A continuación se muestra el diagrama de componentes de las modificaciones propuestas para la implementación de la herramienta Pentaho BI Server 6.1. En la implementación se sugiere utilizar la arquitectura SOA y el *framework* Spring Security. Los servicios serán consumidos por componentes del Pentaho BI Server 6.1 como se muestra en la Figura 26.

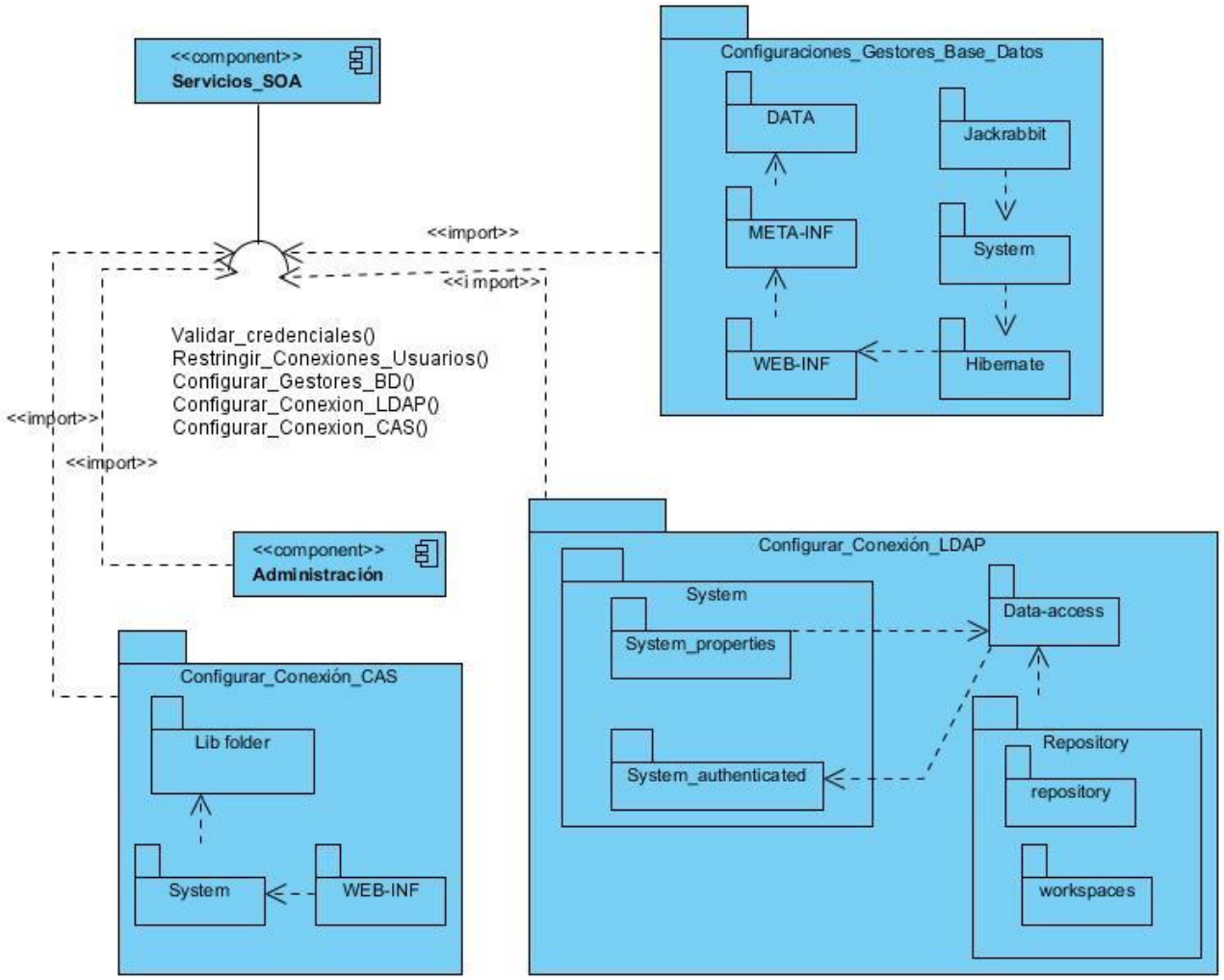


Figura 26: Diagrama de componentes general. Elaboración propia.

Tabla 4: Descripción de componentes. Elaboración propia.

Componente o paquete	Descripción
Servicios_SOA	Componente que contendrá las funcionalidades propuestas. Por medio de una interfaz de programación podrán ser accedidas para ser utilizadas en la implementación.
Administración	Componente que contendrá las clases relacionadas con la autenticación y la verificación de la fortaleza de las contraseñas. Además de las clases relacionadas

	con la implementación del servicio Restringir_Conexiones_Usuarios().
Configuraciones_Gestores_Base_Datos	Paquete que agrupa otros paquetes, componentes y sus relaciones, involucrados en las configuraciones de gestores de base de datos: PostgreSQL, Oracle y MySQL, (Véase, Figura 42 del Anexo 2).
Configurar_Conexión_LDAP	Paquete que agrupa otros paquetes, componentes y sus relaciones, involucrados en las configuraciones para la conexión con un LDAP, (Véase, Figura 43 Anexo 2).
Configurar_Conexión_CAS	Paquete que agrupa otros paquetes, componentes y sus relaciones, involucrados en las configuraciones para adoptar un único proceso de autenticación (CAS), (Véase, Figura 44 del Anexo 2).

La realización de la propuesta de implementación para la herramienta Pentaho BI Server 6.1 se basó en los manuales de configuración que trae la propia versión (help.pentaho.com). Estas configuraciones se deben realizar manualmente, editando varios ficheros de configuración e incluyendo componentes necesarios. Se sugiere partir del diagrama de componentes propuesto y los manuales de ayuda para realizar modificaciones a la herramienta de desarrollo. De esta forma se podrán realizar las operaciones automáticamente, evitando los problemas de seguridad que pueda traer consigo la edición de los parámetros de configuración directamente en los ficheros.

Para realizar estas implementaciones se deberán seguir los 10 principales controles proactivos que propone el proyecto OWASP: verificar la seguridad antes y después, parametrizar las consultas, codificar los datos, validar todas las entradas, implementar controles de identidad y autenticación, implementar controles de acceso, proteger los datos, implementar mecanismos de registros de eventos y detección de intrusos, fortalecer y apoyarse en la seguridad de la tecnología base y gestionar correctamente los errores (González Brito, 2016).

Seguridad. Despliegue de la solución

La comunicación entre los servidores deberá realizarse a través de protocolos seguros como HTTPS. En consecuencia, dicho certificado de seguridad debe incluirse en la máquina virtual de Java, instalada en el servidor de los procesos ETL para la ejecución de los procesos de integración de datos. Este servidor tendrá implementado un *firewall* a nivel de sistema operativo para limitar el acceso de la red perteneciente al servidor de Bases de Datos y Aplicaciones; y para limitar el acceso de la red correspondiente al servidor Fuente de datos (Véase, Figura 27).

La seguridad en los datos también debe garantizarse a través de los elementos físicos que componen la distribución de los componentes de la red, por medio de los métodos de control que poseen, tanto de *hardware* como de *software*. A continuación se presenta una propuesta de despliegue para las soluciones que se desarrollen.

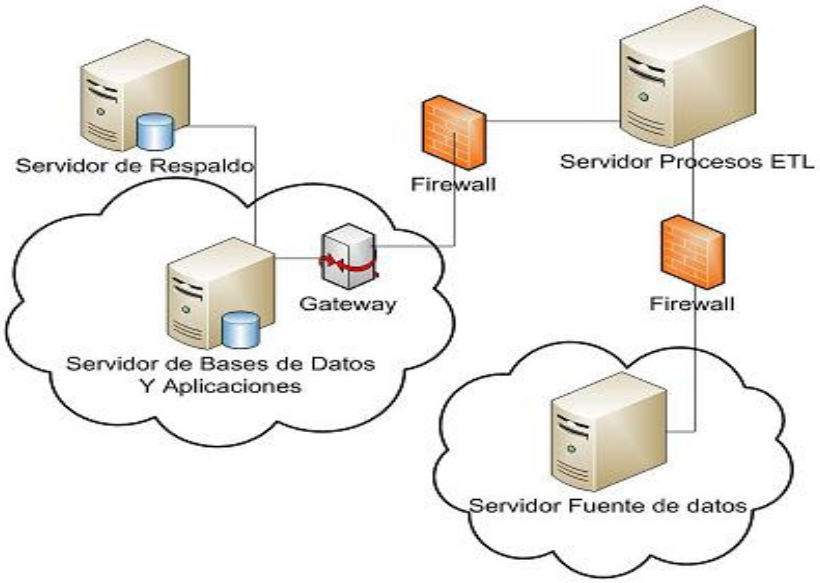


Figura 27: Propuesta de despliegue para la estrategia de seguridad que se propone. Elaboración propia.

2.2.5 Conclusiones parciales

Se propuso una estrategia de seguridad para los almacenes de datos del centro DATEC de la UCI, luego del análisis de las propuestas de seguridad existentes y el diagnóstico de las estrategias de seguridad en este tipo de soluciones. Esta, en específico, parte de los requisitos de seguridad establecidos para cada subsistema.

Se presentaron un conjunto de acciones a seguir, se propusieron principios de seguridad y mecanismos de seguridad para cada subsistema de los AD y se definió una arquitectura de seguridad. Particularmente en el subsistema de visualización de los datos, se propusieron modificaciones a la configuración básica de la herramienta Pentaho BI Server 6.1 para corregir sus restricciones de seguridad.

La Gestión del conocimiento constituyó la herramienta guía para el establecimiento de la estrategia de seguridad propuesta, para lo cual se siguió la metodología de Gestión del conocimiento definida en (Modelo de gestión del conocimiento de la investigación para Colombia y Cuba, 2012).

CAPÍTULO III. APLICACIÓN Y EVALUACIÓN DE LA ESTRATEGIA PROPUESTA

En el presente capítulo se describen algunos de los resultados obtenidos al aplicar la estrategia de seguridad propuesta en un entorno real, en el almacén de datos “Sistema de análisis estadísticos para los procesos eleccionarios en Cuba”. Se realiza la evaluación de la estrategia de seguridad propuesta, para lo cual se utilizó el método Delphi, el cálculo de coeficiente de Kendall y técnicas multicriterios con el consenso de expertos. Se describen cada una de las etapas seguidas en la validación: elección de expertos, elaboración y lanzamiento de cuestionarios, realización de pruebas de concordancia y presentación de los resultados de la evaluación.

3.1 Aplicación de la estrategia de seguridad

La estrategia de seguridad propuesta se aplicó en el almacén de datos “Sistema de análisis estadísticos para los procesos eleccionarios en Cuba”, desarrollado por el centro DATEC de la UCI para la Comisión Nacional Electoral (CEN). Esta solución se inició en el año 2012 y surgió debido a las siguientes problemáticas:

- La información de los procesos electorales a partir del año 2002 hasta 2010 se encontraba en formato Excel.
- A pesar de utilizar la herramienta Microsoft Excel para el almacenamiento de los datos y de las ventajas que posee, los especialistas debían realizar el análisis estadístico de la información de forma manual. Esto podría traer consigo la existencia de errores humanos y la pérdida de información útil para la entidad.
- Se generaban un gran número de datos anuales obstaculizando su análisis. Las fuentes de datos almacenadas en ficheros Excel tenían diferentes formatos, lo que desencadenaba la mala calidad y originaba la no integración de los datos.
- La recuperación y creación de los informes se tornaba engorroso y en ocasiones costoso en cuanto a tiempo y esfuerzo.
- A partir del proceso 2012-2013 la información se encontraba en el sistema operacional desarrollado por la UCI para la gestión de la información.

La Figura 28 muestra su arquitectura general, la cual incluye los diferentes subsistemas (Fuente de datos, Procesos ETL, Almacén de datos, Visualización de los datos) y las herramientas utilizadas en cada uno de ellos.

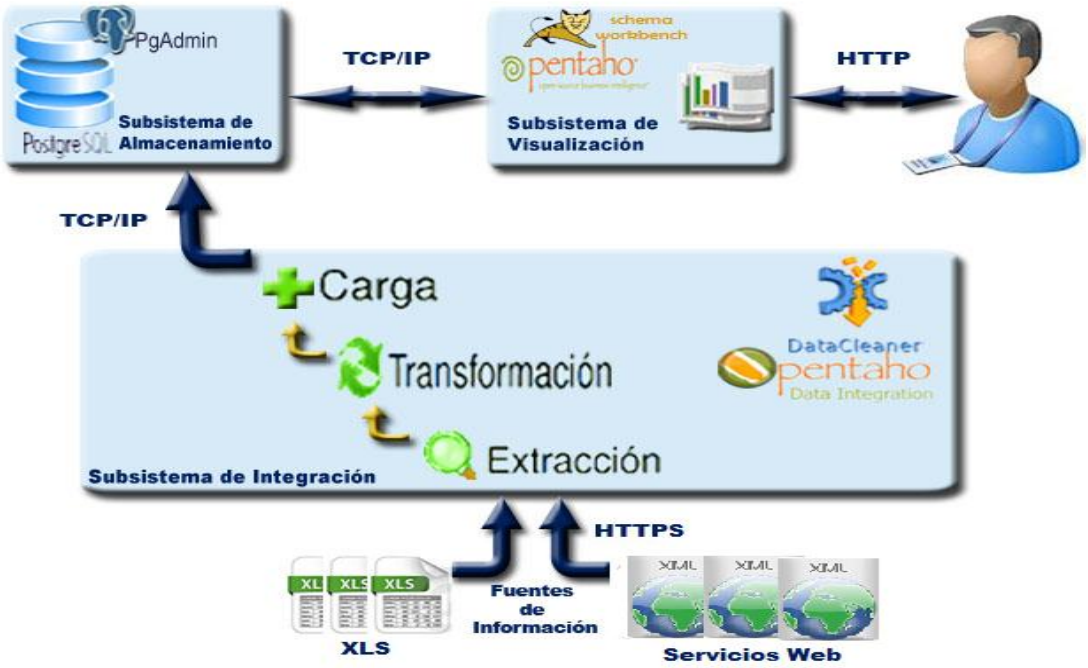


Figura 28: Arquitectura general del almacén de datos. Elaboración propia.

Para realizar la primera etapa de la estrategia de seguridad: acciones a seguir, se partió de los requisitos de seguridad establecidos y se comenzó con la realización del inventario de todos los datos históricos existentes, lo que resultó:

- La mayor cantidad de datos estaba en formato Excel, desde los procesos electorales del año 2002 hasta el año 2010. El volumen de estos datos era aproximadamente de 80 MB.
- Las fuentes de datos en formato Excel tenían diferentes estructuras, lo que ocasionaba la mala calidad de los mismos.
- Se identificaron en esta etapa tres metadatos a utilizar en el proceso de integración, seis nomencladores y cinco conceptos de información que almacenarían datos en el área de almacenamiento intermedio conocida como "*staging area*".

Teniendo en cuenta estos elementos, los datos en el subsistema de almacenamiento se organizaron como ilustra la Figura 29, compuesto por tres esquemas: *elecciones*, *sch_metadatos* y *staging_area*.

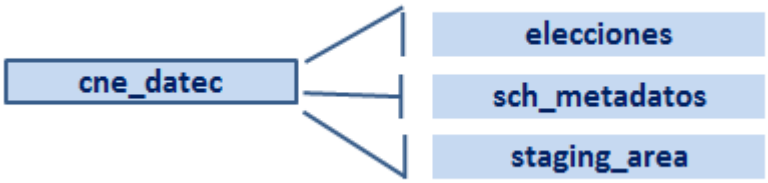


Figura 29: Esquemas del subsistema de almacenamiento. Elaboración propia.

Esquema elecciones: contiene las tablas que almacenan toda la información de los procesos eleccionarios.

Esquema sch_metadatos: contiene las tablas de metadatos utilizadas en el proceso de integración de datos por los especialistas de ETL.

Esquema staging_area: contiene tablas temporales necesarias por los especialistas de ETL en el proceso de integración de datos.

Al llevar a cabo la acción de clasificación de los datos, se clasificaron todos los datos en secretos por ser estrictamente sensibles para la CEN.

Para la seguridad en la BD se definió el rol Administrador de BD el cual posee total acceso a la BD, además tiene la responsabilidad de llevar a cabo la política de respaldo y recuperación. De igual manera, se definió el rol Administrador de ETL, el cual tiene permisos de lectura-escritura sobre el esquema elecciones y todos los permisos sobre los esquemas *sch_metadatos* y *staging_area*, pues es el encargado de realizar los procesos de integración. Para la seguridad en el subsistema de visualización se definieron los roles Administrador y Analista, el primero tiene acceso total al área de análisis que incluye todos los libros de trabajos de los reportes y administra los usuarios, roles y reportes. El segundo tiene acceso de solo lectura a dicha área de análisis (Véase, Tabla 5).

Tabla 5: Roles y permisos. Elaboración propia.

Roles	Permisos
Administrador de base de datos	Total acceso a la base de datos. Lleva a cabo la política de respaldo y recuperación.
Administrador de ETL	Permiso de lectura y escritura sobre el esquema elecciones y todos los permisos sobre los esquemas *sch_metadatos* y *staging area*.
Administrador	Acceso total al área de análisis a nivel de aplicación de visualización. Administra los usuarios, roles y reportes.
Analista	Tiene acceso de solo lectura a las diferentes áreas de análisis, que incluyen los libros de trabajo de los reportes de la aplicación.

En las sucesivas acciones se realizó lo siguiente:

Identificación de las vulnerabilidades en la seguridad de los datos: las vulnerabilidades que se identificaron estuvieron relacionadas con la disponibilidad, confidencialidad e integridad de la información.

Identificación de medidas de protección de datos: las vulnerabilidades identificadas en la etapa anterior permitieron identificar los mecanismos de seguridad para cada subsistema de la arquitectura propuesta en la presente investigación; ajustada a las características de los proyectos de almacenes de datos del centro DATEC y basada en los principios de seguridad propuestos.

Ejemplo de estas medidas fueron:

Para proteger la fuente de datos se aplicaron mecanismos de seguridad desde el sistema operativo: *SELinux* para las fuentes de datos pertenecientes a los procesos electorales del año 2002 hasta el año 2010 que se encontraban en formato de archivos Excel.

Por otro lado, para proteger la información de los procesos electorales 2012-2013, se estableció un esquema de seguridad protegiendo los objetos de la bases de datos con roles y privilegios. Se controló además la autenticación de clientes al sistema gestor de bases de datos PostgreSQL mediante la manipulación de los parámetros del fichero pg_hba.conf de la siguiente manera (Véase, Figura 30):

# TYPE	DATABASE	USER	CIDR-ADDRESS	METHOD
host	cne_datec	ALL	0.0.0.0/0	reject
host	cne_datec	admin_db, admin_etl	192.168.2.2/32	md5

Figura 30: Autenticación de clientes al sistema gestor de bases de datos. Elaboración propia.

Como se evidencia en la configuración anterior de los parámetros del fichero pg_hba.conf, tendrán acceso a la base de datos cne_datec los roles admin_db, admin_etl que se conecten desde el servidor de ETL con dirección IP fija suministrando una contraseña encriptada con md5, excluyendo al resto de las conexiones desde internet.

En el diseño y ejecución de los procesos de integración de datos utilizando la herramienta Pentaho Data Integration 6.1, se implementaron los siguientes mecanismos de seguridad:

Seguridad basada en el repositorio de ficheros de la herramienta Pentaho Data Integration 6.1 para los ficheros planos almacenados en el disco duro (Véase, Figura 31).

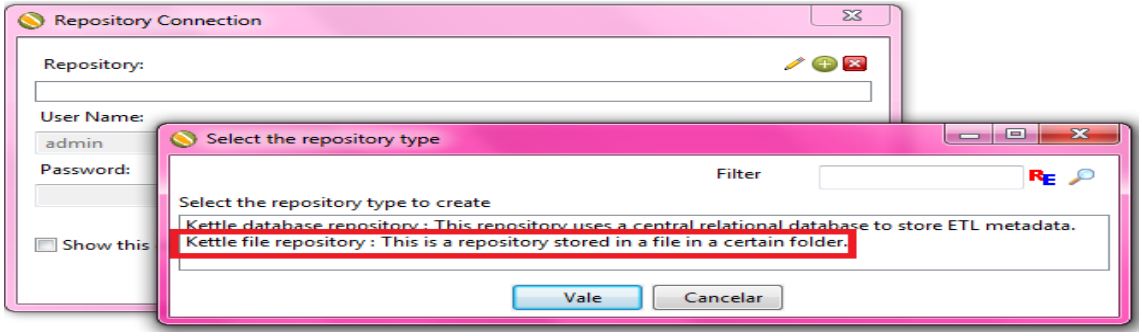


Figura 31: Opciones de repositorios del PDI. Elaboración propia.

Seguridad basada en el repositorio de bases de datos de la herramienta Pentaho Data Integration 6.1 para la fuente de datos perteneciente a la base de datos del sistema operacional (Véase, Figura 32).

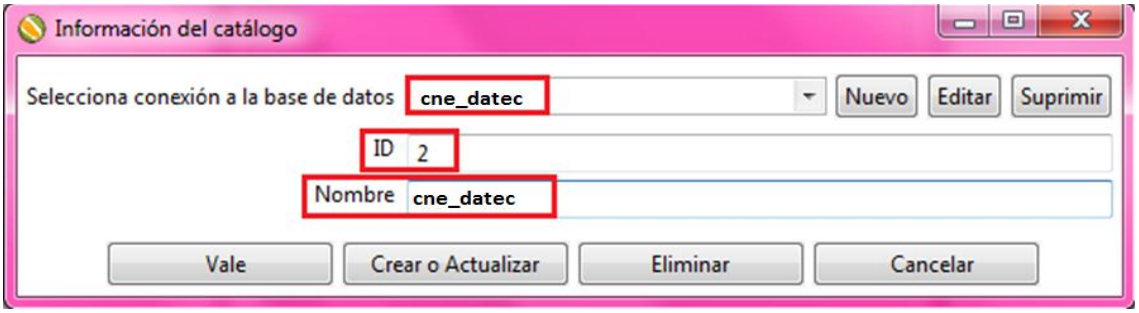


Figura 32: Ejemplo de repositorio de base de datos. Elaboración propia.

La seguridad basada en el repositorio de ficheros y bases de datos de la herramienta Pentaho Data Integration 6.1 permite la selección y el trabajo con los objetos del repositorio mediante sistema de autenticación de usuarios (Véase, Figura 33).

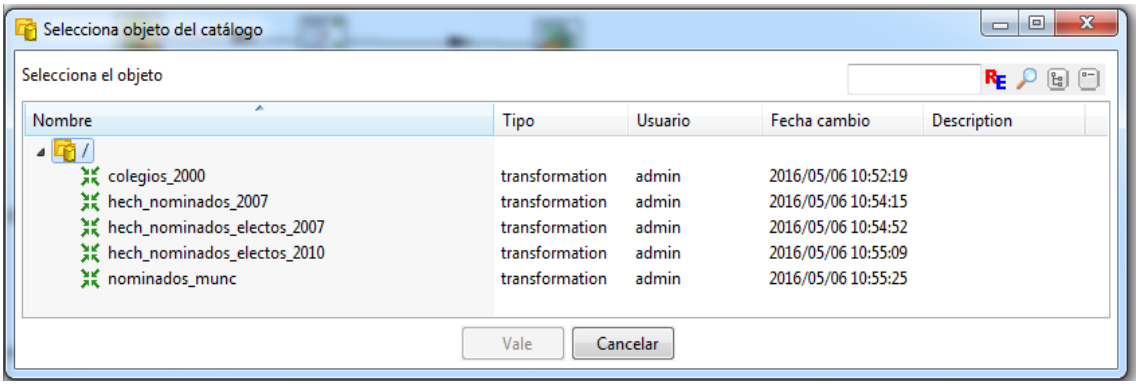


Figura 33: Ejemplo de trabajo con repositorios. Elaboración propia.

Para el intercambio de información entre el subsistema Fuente de datos y el servidor de Procesos ETL, se utilizó el cálculo del *hash* de cada fichero para verificar la integridad de la fuente de información. Este fue comprobado haciendo uso de la herramienta de integración Pentaho Data Integration 6.1, como se muestra a continuación:

```
//Cálculo del hash de ficheros.

file = new Packages.java.io.File(filename.getString());
fileInputStream = new Packages.java.io.FileInputStream(file);
var content = Packages.org.pentaho.di.core.Const.createByteArray(file.length());
var md5_hash = org.apache.commons.codec.digest.DigestUtils.md5Hex(content);
fileInputStream.close();


//Comprobación del hash enviado.

var arg1= indexOf(short_filename.getString(),"_") + 1;
var arg2= indexOf(short_filename.getString(),".") - arg1;

md5_hash_previous = substr(short_filename.getString(),arg1,arg2);


var result_hash;
if(md5_hash == md5_hash_previous)
   result_hash = true;
else
     result_hash = false;
```

Figura 34: Comprobación de *hash* con componente *Java Script* de la herramienta Pentaho Data Integration 6.1. Elaboración propia.

Además, se incluyó el certificado de seguridad del sistema operacional a la máquina virtual de Java en el servidor Procesos ETL, lo cual permitió una comunicación segura entre la herramienta PDI y la fuente de datos del sistema operacional. De esta manera la comunicación entre ambos fue a través del protocolo HTTPS.

Otro ejemplo de mecanismo de seguridad utilizado en la implementación de los procesos de integración fue el almacenamiento de los metadatos utilizando la codificación *Base64* que incorpora la herramienta PDI, como se muestra en la presente figura:

```
//Implementación del algoritmo de encriptación Base64

var bytes = Packages.org.apache.commons.codec.binary.Base64.decodeBase64(F1.getString().getBytes() );
var decString = new Packages.java.lang.String( bytes );

var encString = new Packages.java.lang.String( Packages.org.apache.commons.codec.binary.Base64.encodeBase64
( decString.getBytes() ) );


Alert(decString);

Alert(encString);
```

Figura 35: Implementación del algoritmo Base64. Elaboración propia.

Esta implementación utilizando la codificación *Base64* fue realizada en el componente Java Script en la transformación “Cargar Metadatos”, de todos los trabajos implementados en la solución (Véase, Figura 36).

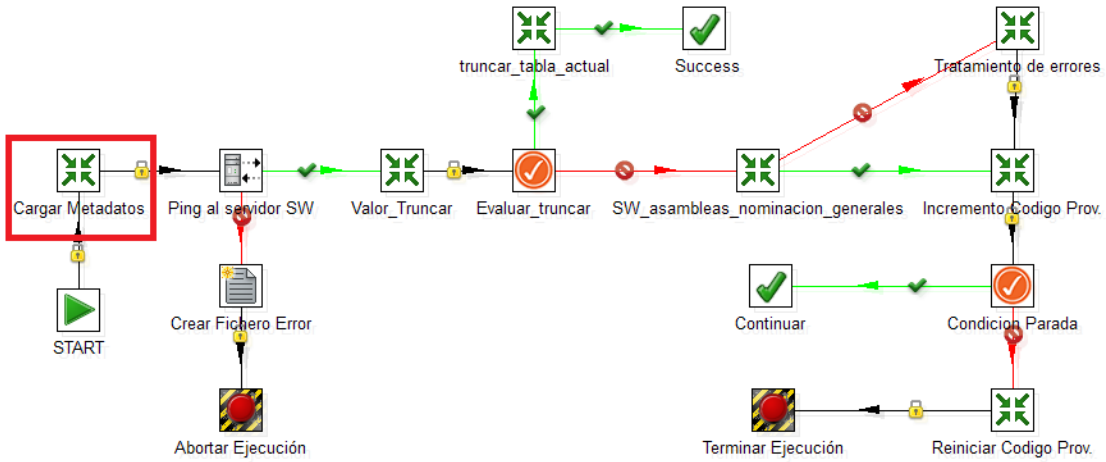


Figura 36: Utilización de la codificación *Base64* en los procesos de integración. Elaboración propia.

En la base de datos del AD se identificaron los tres sectores de la arquitectura establecida: **Integrado**, **Línea Cerca** y **Archivado**. En el sector **Integrado** se almacenaron los datos del proceso electoral actual o último proceso electoral, por ser los datos de mayor acceso según el objetivo del proyecto. En el mismo, el volumen de información fue poco y la respuesta del sistema ante la visualización de reportes estadísticos actuales por los analistas fue de pocos segundos. Para separar la información del proceso eleccionario actual o último de los demás procesos en el sector **Integrado**, se utilizó el particionado de tablas. Esto se implementó mediante la herencia entre tablas que ofrece el gestor de base de datos PostgreSQL utilizado en la implementación, lo cual permitió las siguientes ventajas**:**

- Reducir la cantidad de datos a recorrer en cada consulta SQL y posteriormente en las consultas MDX implementadas en los reportes de la aplicación de inteligencia de negocios.
- Aumentar el rendimiento, pues menos datos que recorrer trae consigo una ejecución más rápida.

Por otro lado, si se necesitan realizar comparaciones históricas entre los datos de diferentes procesos se podrán realizar las consultas sobre las tablas padres, las cuales contendrán toda la

información debido a su relación de herencia. A continuación se muestra la implementación de herencia entre las tablas hech_calidad_voto_actual y hech_calidad_voto:

```
---Cambiando al rol admin_db
SET SESSION AUTHORIZATION admin_db;

---Herencia entre hech_calidad_voto_actual y hech_calidad_voto
CREATE TABLE elecciones.hech_calidad_voto_actual
(
  dim_dpa_id integer NOT NULL,
  dim_tipo_voleta_id integer NOT NULL,
  dim_num_vuelta_id integer NOT NULL,
  dim_procesos_id integer NOT NULL,
  boletas_blanco integer NOT NULL,
  boletas_anuladas integer NOT NULL,
  boletas_validas integer NOT NULL,
  votos_unidos integer NOT NULL,
  votos_selectivos integer NOT NULL,
  circunscripciones_prox_vuelta integer,
  num_prox_vuelta_id integer,
  CONSTRAINT "Refdim_dpa67" FOREIGN KEY (dim_dpa_id)
      REFERENCES elecciones.dim_dpa (dim_dpa_id) MATCH SIMPLE
      ON UPDATE NO ACTION ON DELETE NO ACTION,
  ---demás atributos
)INHERITS ("elecciones"."hech_calidad_voto")
```

Figura 37: Ejemplo de implementación de herencia entre tablas. Elaboración propia.

El sector **Línea Cerca** no se utilizó pues todos los datos del sector **Integrado** tenían la misma probabilidad de acceso y no existió gran volumen de información. Cabe destacar según se explicó anteriormente, que la implementación de este sector se realiza de manera opcional, pues generalmente se introduce cuando existe un gran volumen de información en el sector **Integrado** y se trasladan a él los datos con una baja probabilidad de acceso.

La distribución física de los elementos de *hardware* necesarios que se utilizó fue la especificada en la Figura 27. De igual forma, teniendo en cuenta otras características y vulnerabilidades se establecieron otros mecanismos de seguridad. En el sistema de visualización se utilizó la herramienta Pentaho BI Server en su versión 6.1, a la cual se le sugieren modificaciones en el capítulo 2 de la presente investigación. Finalmente se garantizó la seguridad en cada subsistema del AD, lo cual se puede corroborar considerando que desde el año 2012 hasta la actualidad no se han detectado incidentes de seguridad.

3.2 Métodos seleccionados para validar la propuesta

Teniendo en cuenta que para propuestas de este tipo la recolección de datos posterior a su aplicación puede demorar mucho tiempo, y en este caso aún no se tiene suficiente información sobre resultados obtenidos al aplicar la propuesta, fueron seleccionados: el método Delphi, el cálculo del coeficiente de Kendall para realizar las pruebas de concordancia y técnicas multicriterios para evaluar el índice de impacto de la estrategia de seguridad.

El método Delphi se clasifica como uno de los métodos generales de prospectiva, que busca acercarse al consenso de un grupo de expertos con base en el análisis y la reflexión de un problema definido. Consiste en la selección de un grupo de expertos a los que se les solicita su opinión sobre cuestiones referidas a acontecimientos del futuro. Las estimaciones de los expertos se realizan en sucesivas rondas, anónimas, con el objetivo de lograr el consenso, pero con la máxima autonomía por parte de los participantes. Para su aplicación es necesario considerar los siguientes aspectos, que constituyen características básicas del método:

- **Retroalimentación controlada**: antes del comienzo de una nueva ronda los expertos conocen los resultados alcanzados en la ronda anterior para recapacitar sobre las respuestas que se ofrecen en la ronda precedente.
- **Proceso interactivo**: se realizan cuestionarios sucesivos para disminuir la dispersión de las opiniones. En la investigación se realizaron dos rondas de aplicación del cuestionario a los expertos seleccionados.
- **Anonimato**: los miembros del grupo no conocen a quien corresponden las respuestas que analizan, aunque sí conocen quiénes son los demás expertos (se utilizó una lista de correo electrónico en la que aparecían los nombres de todos). Las respuestas se presentan de manera global y conjunta, por lo que no pueden conocer lo que opina cada uno de ellos. Este anonimato, lo promovió la autora de la investigación, porque coincide con (Gilson, y otros, 2009) en que esta forma promueve la libertad para expresar la opinión personal.

3.3 Etapas presentes en la validación

Las actividades para la validación de la estrategia de seguridad se organizan en cuatro etapas fundamentales, las cuales se pueden representar en dos fases como se muestra en la Figura 38.

Las etapas son:

- Elección de expertos.
- Elaboración y lanzamiento de los cuestionarios.
- Realización de pruebas de concordancia mediante el cálculo del coeficiente de Kendall.
- Presentación de los resultados de la evaluación.

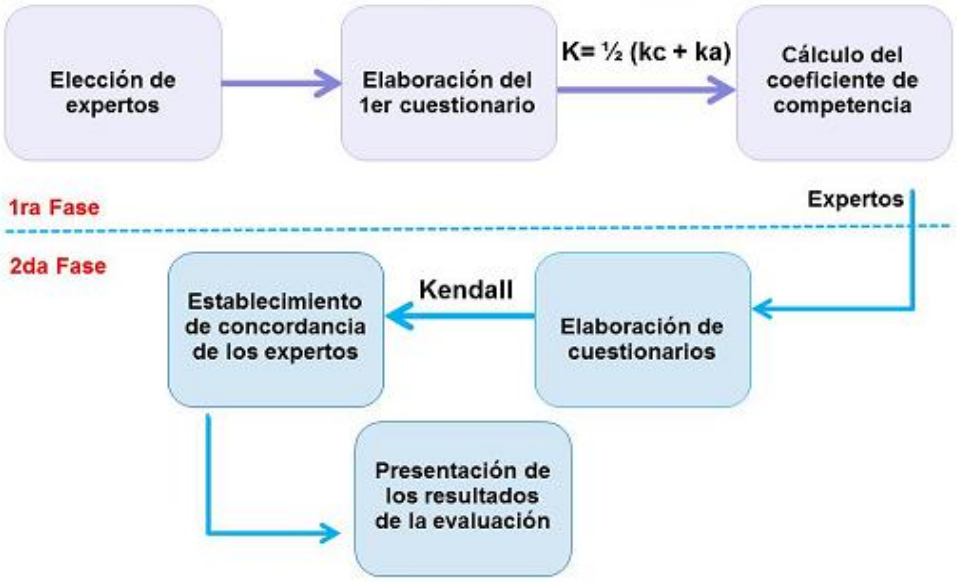


Figura 38: Etapas utilizadas en la validación de la estrategia de seguridad. Elaboración propia.

3.3.1 Elección de expertos

Al aplicar el método Delphi no existe una norma para identificar el número determinado de expertos a participar en el proceso de validación. Se sugiere que participen entre siete y 30 expertos, considerando que con siete la probabilidad de error disminuye exponencialmente y con más de 30 aunque continúa disminuyendo, es poco significativo y no compensa el incremento de costo y esfuerzo.

En la etapa de selección de expertos se identificaron 10 especialistas, a los cuales se les entregó o envió en una primera fase un cuestionario solicitando su conformidad para ser experto en la validación de la estrategia de seguridad propuesta. Se les solicitó la labor que realiza y su calificación profesional. Además, se solicitó una autoevaluación relacionada con el grado de conocimiento del experto acerca del tema de investigación y las fuentes a partir de las cuales han logrado el conocimiento, estableciendo su grado de influencia (Véase, Anexo 3).

Al obtener los resultados de la primera encuesta aplicada al panel de expertos, se procedió con el cálculo de coeficiente de competencia. Este coeficiente se determina a través de la fórmula $\boldsymbol{K = \frac{1}{2}(Kc + Ka)}$.

Donde:

Kc es coeficiente de conocimientos y Ka es el coeficiente de argumentación. El coeficiente Kc se obtiene de la primera tabla de la encuesta, la cual indica la autoevaluación del grado de conocimiento del experto. El experto indica marcando la casilla enumerada en una escala del 1 al 10, interpretándose 1 como no tener ningún conocimiento y 10 el pleno conocimiento de la problemática tratada. Posteriormente para ajustarla a la teoría de las probabilidades se multiplica por 0.1 el valor de la casilla seleccionada. El coeficiente Ka se calcula utilizando las respuestas relacionadas a la segunda tabla del cuestionario. Este coeficiente se autoevalúa en alto (A), medio (M) o bajo (B) para un grupo de fuentes que influyen sobre el nivel de fundamentación del tema de la investigación.

Tabla 6: Resultados de procesamiento para la determinación del coeficiente de competencia

Experto	Kc	Ka	K	Valoración
1	0.8	0.9	0.85	A
2	0.9	0.9	0.9	A
3	0.9	0.9	0.9	A
4	1	1	1	A
5	0.9	1	0.95	A
6	0.7	1	0.85	A
7	0.9	0.9	0.9	A
8	0.9	0.9	0.9	A
9	0.8	0.9	0.85	A
10	0.7	0.9	0.8	A

La valoración para la selección de los expertos se realizó teniendo en cuenta el código de interpretación de tales coeficientes de competencias (Véase, Tabla 6) de la siguiente forma:

- Si 0,8 < K < 1,0 coeficiente de competencia **alto**.
- Si 0,5 < K < 0,8 coeficiente de competencia **medio**.
- Si K < 0,5 coeficiente de competencia **bajo**.

Se seleccionaron todos los expertos consultados debido a que el coeficiente de todos resultó ser alto. Además, teniendo en cuenta su grado científico (Véase, Figura 39), los años de experiencia en el desarrollo de software y en la seguridad de aplicaciones (Véanse, Figura 40 y Figura 41).

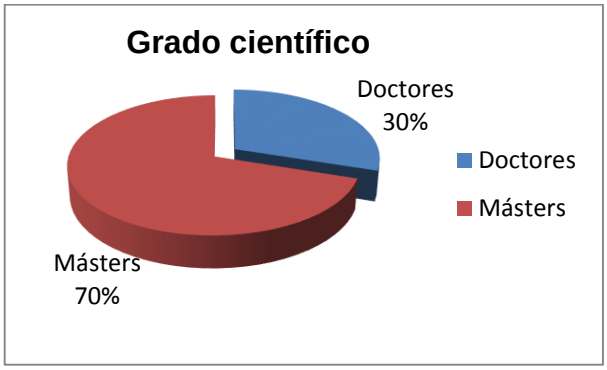


Figura 39: Grado científico del panel de expertos. Elaboración propia.

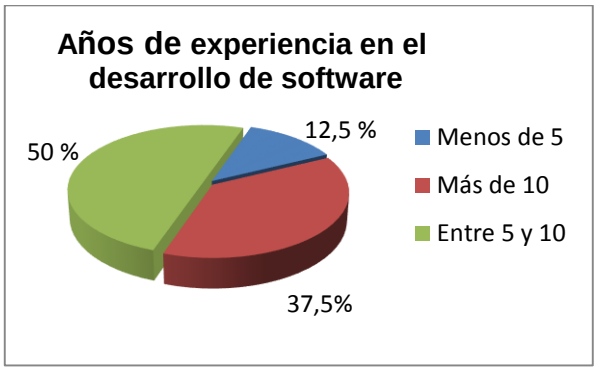


Figura 40: Años de experiencia del panel de expertos en el desarrollo del software. Elaboración propia.

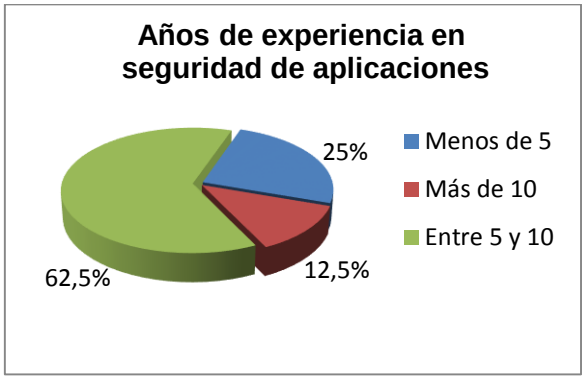


Figura 41: Años de experiencia del panel de expertos en la seguridad de aplicaciones. Elaboración propia.

3.3.2 Elaboración y lanzamiento de cuestionarios

En la elaboración de los cuestionarios para la validación se propusieron a los expertos un grupo de criterios agrupados en categorías, los cuales podían ser ratificados como medidor de la propuesta o ser modificados (Véase, Anexo 4). Si el experto consideraba que el criterio estaba mal categorizado podía realizar una nueva propuesta, de igual manera si consideraba que debía incluir otros criterios. Se solicitó una propuesta de peso para las categorías y una evaluación de cada criterio. Las respuestas debían estar enmarcadas en una escala numérica representando a las calificaciones: Muy bueno, Bueno, Regular, Pobre y Malo. Todos los criterios y categorías fueron ratificados por los 10 expertos en una segunda ronda. Los expertos también emitieron su criterio relacionado a la importancia que tiene una categoría respecto al resto de las categorías

(Véase, Anexo 5). Para que los especialistas pudieran responder estos cuestionarios se les envió un resumen de la estrategia de seguridad desarrollada.

3.3.3 Realización de pruebas de concordancia

Para que la propuesta tenga una mayor validez es necesario que exista un adecuado acuerdo entre los integrantes del panel de expertos. Esto se comprueba mediante el cálculo del coeficiente de concordancia de Kendall, el cual ayuda a precisar el grado de coincidencia de las valoraciones realizadas por los expertos. Para determinar este coeficiente se construye una tabla de aspectos a evaluar contra expertos. Estos datos fueron obtenidos al aplicar las encuestas realizadas en la segunda fase. Luego de haber confeccionado dicha tabla se siguen algunos pasos para calcular Chi Cuadrado real ($\boldsymbol{X^2_{real}}$).

Dados n el número total de criterios (13 criterios) a evaluarse y m el número de expertos (10 expertos) involucrados se realizan los siguientes pasos para determinar la consistencia del trabajo de los expertos:

Calcular para cada criterio la sumatoria del peso dado por cada experto mediante la expresión: $\sum \boldsymbol{E}$

Determinar el valor de puntuación promedio de cada criterio: $\boldsymbol{EP}$

Se calcula el peso medio dado por cada experto $\boldsymbol{M \sum E}$ y luego la diferencia $\boldsymbol{\Delta C = \sum E - M \sum E}$

Determinar la desviación de la media y elevar el resultado al cuadrado para obtener la dispersión (S), dada por la expresión: $\boldsymbol{S = \sum (\Delta C)^2}$

La Tabla 16 del Anexo 6 muestra los pesos otorgados por los 10 expertos a los 13 criterios planteados, a partir de los cuales se calculó el valor de la dispersión como se representa en la Tabla 7.

Tabla 7: Cálculo de dispersión. Elaboración propia.

Criterios	∑E	$\boldsymbol{EP}$	$\boldsymbol{\Delta C}$	$\boldsymbol{(\Delta C)^2}$
C1	48	4.80	1.23	1.51
C2	43	4.30	3.77	14.21
C3	50	5	3.23	10.43
C4	49	4.90	2.23	4.97
C5	48	4.80	1.23	1.51
C6	48	4.80	1.23	1.51
C7	49	4.90	2.23	4.97
C8	48	4.80	1.23	1.51
C9	46	4.60	0.77	0.59
C10	46	4.60	0.77	0.59
C11	40	4.00	6.77	45.83
C12	44	4.40	2.77	7.67
C13	49	4.90	2.23	4.97
M∑E	46.77			

S= $\sum(\Delta \boldsymbol{C})^{\boldsymbol{2}}$	100.27

Conociendo la dispersión (S) se puede calcular el coeficiente de concordancia de Kendall dado por la expresión: $\boldsymbol{K} = \frac{\boldsymbol{S}}{\left(\boldsymbol{m^2(n^3-n)}\right)/\boldsymbol{n}}$

$$\boldsymbol{K} = \frac{\boldsymbol{100.27}}{\frac{\boldsymbol{10^2(13^3-13)}}{\boldsymbol{13}}}$$

$$\boldsymbol{K} = \boldsymbol{100.27/16800}$$

$$\boldsymbol{K} = \boldsymbol{0.0060}$$

El coeficiente de Kendall permite calcular el Chi cuadrado real a partir de la siguiente expresión:

$$\boldsymbol{X^2_{real}} = \boldsymbol{m(n-1)K}$$

$$\boldsymbol{X^2_{real}} = \boldsymbol{10(13-1)0.0060}$$

$$\boldsymbol{X^2_{real}} = \boldsymbol{0.72}$$

Se compara el Chi cuadrado real calculado con el que se obtiene de la Tabla de Distribución Chi Cuadrado, se toma 1 – α = 0.95 donde α = 0.05 es el error permisible. Si se cumple que $\boldsymbol{X^2_{real}} < \boldsymbol{X^2_{tabla}}$ puede decirse que existe concordancia en el trabajo de los expertos (Siegel, 1974).

El Chi cuadrado obtenido se comparó con los valores críticos de las tablas de distribución para α= 0.05 y 12 grados de libertad (n-1), $\boldsymbol{X^2_{tabla}} = \boldsymbol{21.03}$

Al cumplirse que **0.72 < 21.03**, se llega a la conclusión de que existe concordancia entre los expertos y no es necesario realizar otra iteración.

3.3.4 Resultados de la evaluación

Para el cálculo del índice de impacto de la estrategia de seguridad se utilizaron técnicas multicriterios (Barba Romero, 1987), similar a su aplicación en (Sistema informático para calcular el Índice de Control Interno en instituciones cubanas, 2016) y en (Método multicriterio multiexperto para evaluar el impacto de las TICs en la formación del ingeniero en Ciencias Informáticas, 2016).

Conociendo el peso relativo de cada categoría y la calificación cuantitativa dada por los expertos en una escala de 1 a 5 (Malo, Pobre, Regular, Bueno y Muy bueno), se determinó el índice de impacto (II) de la estrategia de seguridad propuesta, dado por la expresión: $\boldsymbol{II} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{P_i Cp_i}$

Donde:

$\boldsymbol{P_i}$: Peso relativo de cada categoría, valor que atribuye la importancia que tiene una categoría respecto al resto de las categorías. La sumatoria de estos valores relativos debe ser igual a 100. Se calcula a través de la fórmula: $\boldsymbol{P_i} = \frac{\boldsymbol{EP_i}}{\boldsymbol{100}}$

$\boldsymbol{C_{pi}}$: Calificación cuantitativa dada por los expertos a cada categoría en una escala numérica de 1 a 5.

La Tabla 8 visualiza el resultado de la encuesta realizada (Véase, Anexo 6) donde es posible determinar el peso de importancia de cada categoría respecto al resto (Véanse los valores $\boldsymbol{P_i}$).

Tabla 8: Asignación de peso a las categorías. Elaboración propia.

Expertos	Categoría 1	Categoría 2	Categoría 3	Categoría 4	Categoría 5
E1	15	25	25	10	25
E2	30	15	15	10	30
E3	20	10	20	20	30
E4	20	40	20	10	10
E5	25	20	20	10	25
E6	15	35	25	10	15
E7	15	25	20	15	25
E8	20	25	25	10	20
E9	20	25	20	15	20
E10	20	40	20	10	10
$\boldsymbol{EP}$	20	26	21	12	21
$\boldsymbol{P_i}$	0.20	0.26	0.21	0.12	0.21

A continuación se muestra la tabla categoría contra evaluación emitida, donde C1…C5 son los categorías evaluadas y E1…E5 la evaluación de los expertos respecto a 5.

Tabla 9: Determinación del índice de impacto. Elaboración propia.

Expertos	Categoría 1	Categoría 2	Categoría 3	Categoría 4	Categoría 5
E1	5	5	5	5	5
E2	5	5	5	3	4
E3	5	5	4	5	5
E4	5	4	4	4	5
E5	4	5	5	5	5
E6	4	5	5	4	4
E7	4	5	5	5	5
E8	5	5	4	5	5
E9	5	4	5	4	5
E10	4	5	5	4	5
$\boldsymbol{EP}$	4.60	4.80	4.70	4.30	4.80
$\boldsymbol{C_{pi} = \frac{EP}{5}}$	0.92	0.96	0.94	0.86	0.96
$\boldsymbol{P_i}$	0.20	0.26	0.21	0.12	0.21
$\boldsymbol{P_i Cp_i}$	0.18	0.25	0.20	0.10	0.20
$\boldsymbol{\sum P_i Cp_i}$	**0.93**				

El índice de impacto (II = 0.93) es **alto** por lo que se puede afirmar que existe una alta probabilidad de éxito al aplicar la estrategia de seguridad, teniendo en cuenta que:

II> 0.7: **Alta** probabilidad de éxito

0.7 > II > 0.5: Probabilidad **media** de éxito

0.5 > II > 0.3: Probabilidad de éxito **baja**

0.3 > II: **Fracaso** seguro

3.4 Conclusiones parciales

La aplicación de la “Estrategia de seguridad para soluciones de Almacenes de datos” en el desarrollo del almacén de datos “Sistema de análisis estadísticos para los procesos eleccionarios en Cuba” permitió elevar la seguridad en cada subsistema presente, lo cual puede corroborarse considerando que desde su desarrollo en el año 2012 hasta la actualidad no se han detectado incidentes de seguridad.

Como parte del proceso de validación se realizaron consultas a expertos que emitieron valoraciones favorables sobre la estrategia de seguridad. Los expertos avalaron las categorías y los criterios propuestos para realizar la validación, existiendo un adecuado acuerdo entre los mismos, lo que se demostró a través del cálculo de coeficiente de Kendall. Finalmente la aplicación de técnicas multicriterios permitió evaluar el índice de impacto de la estrategia de seguridad propuesta, obteniéndose resultados que califican a la solución con una alta probabilidad de éxito.

CONCLUSIONES

El estudio relacionado con el proceso de implementación de mecanismos de seguridad en diferentes arquitecturas de almacenes de datos, permitió conocer que existen propuestas de inclusión de la seguridad para estas soluciones; pero inconvenientemente están dirigidas a elementos específicos presentes en su construcción.

Específicamente en el centro DATEC de la UCI, la metodología utilizada para el desarrollo de almacenes de datos especifica las etapas en las que deben implementarse los requisitos relacionados con la confidencialidad, disponibilidad e integridad de la información. Inconvenientemente, no se encontró una estrategia para elevar la seguridad en todo el ciclo de desarrollo de estas soluciones.

Como resultado de la presente investigación se elaboró una estrategia de seguridad para soluciones de almacenes de datos del centro DATEC de la UCI. La estrategia de seguridad propuesta incluye una guía de acciones para los desarrolladores con el propósito de lograr sistemas seguros desde los subsistemas que se desarrollan. Las acciones organizadas en una espiral de mejora continua contribuyen a elevar la calidad de dichas aplicaciones a través de su seguridad. Además, la estrategia de seguridad aborda principios de seguridad, mecanismos de seguridad y una propuesta de arquitectura de seguridad.

Se aplicó la estrategia de seguridad en el almacén de datos: “Sistema de análisis estadísticos para los procesos eleccionarios en Cuba”; donde se emplearon buenas prácticas de seguridad en cada subsistema presente. Finalmente, se evaluó la factibilidad de aplicar la estrategia de seguridad utilizando técnicas multicriterios con el consenso de expertos, cuyos resultados califican la propuesta con una alta probabilidad de éxito.

RECOMENDACIONES

Al concluir la presente investigación, se recomienda para el desarrollo de estudios futuros:

- Extender la estrategia de seguridad propuesta para soluciones de almacenes de datos del centro DATEC de la UCI a otros entornos donde se desarrollen estas soluciones.
- Realizar las modificaciones necesarias a la herramienta Pentaho BI Server 6.1, partiendo de los diagramas de componentes que se proponen en la investigación.
- Incorporar nuevas propuestas de mecanismos de seguridad que surjan de experiencias de etapas de trabajo anteriores.

REFERENCIAS BIBLIOGRÁFICAS

1. **27001, ISO/IEC. 2005.** *Information technology - Security techniques - Information.* s.l. : International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), 2005.
2. *A Schematic Technique Using Data type Preserving Encryption to Boost Data Warehouse Security.* **M.Sreedhar, Reddy , y otros. 2011.** India : s.n., 2011.
3. *A survey on current security perspectives in data warehouses.* **G., Thangaraju y X.Agnes , Kala Rani. 2016.** 0976, Enero de 2016, International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), Vol. 19.
4. *A UML 2.0/OCL Extension for Designing Secure Data Wharehouses.* **Villarroel, Rodolfo, y otros. 2006.** 1, febrero de 2006, Journal of Research and Practice in Information Technology, Vol. 38.
5. *Access control and audit model for the multidimensional modeling of data warehouses.* **Villarroel, Rodolfo, y otros. 2006.** 2006, Decision Support Systems 42.
6. **Aguilar Mayorga, Sandra Mireya y Lemus Castiblanco, Jorge Leonardo. 2009.** *Pentaho - BI.* 2009.
7. *An Improved Security Framework for Data Warehouse: A Hybrid Approach.* **Ahmad, S y Ahmad, R. 2010.** s.l. : IEEE, 2010.
8. *Applying an MDA-based approach to consider security rules in the development of secure DWs.* **Blanco, Carlos. 2009.** 2009. nternational Conference on Availability, Reliability and Security.
9. *Arquitecturas de Nueva Generación .* **Correal, Dario. 2016.** 2016.
10. **Barba Romero, Sergio. 1987.** *Panoramica actual de la desición multicriterio discreta.* Universidad de Alcalá de Henares : s.n., 1987.
11. **Barranco Fragoso, Ricardo. 2012.** https://www.ibm.com. [En línea] 18 de 06 de 2012. https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/.
12. *Basic concepts and taxonomy of dependable and secure computing.* **Avizienis, A, y otros. 2004.** 2004.
13. **Berzal Galiano, Fernando. 2016.** Bases de Datos. Fundamentos de Diseño de Bases de Datos. [En línea] 2016. http://elvex.ugr.es/idbis/db/docs/dw.pdf.
14. **Bouman, Roland y Dongen, Jos van. 2009.** *Pentaho Solutions, Business Intelligence and Data Warehousing with Pentaho and MySQL.* Indianapolis : Wiley Publishing, Inc. 10475 Crosspoint Boulevard, 2009. IN 46256.
15. *Building a secure star schema in data warehouse by an extension of rela-tional package from CWM.* **Soler, Emilio. 2008.** 2008.
16. **Calidad & Gestión. Consultoría para empresas. 2015.** http://www.calidad-gestion.com.ar. *http://www.calidad-gestion.com.ar.* [En línea] Buenos Aires - Argentina, 2015. http://www.calidad-gestion.com.ar/boletin/58_ciclo_pdca_estrategia_para_mejora_continua.html.

17. **CAS overview. 2016.** Kungliga Tekniska Hogskolan. [En línea] 2016. https://www.kth.se/social/group/cas/page/cas-overview/.

18. **Cryptome. 2013.** Fundamental Security Concepts. [En línea] 2013. https://cryptome.org/2013/09/infosecurity-cert.pdf.

19. **Data Warehouse Security Considerations. 2016.** Data Warehouse Security Considerations. [En línea] 2016. [Citado el: 07 de Junio de 2016.] https://sbri.innovateuk.org/c/document_library/.

20. **Data Warehousing - Security. 2016.** Tutorialspoint. [En línea] 2016. [Citado el: 31 de Mayo de 2016.] http://www.tutorialspoint.com/dwh/dwh_security.htm.

21. *Data Warehousing 2.0 and SQL Server: Architecture and Vision.* **Inmon, Bill. 2009.** 2009.

22. *Data Warehousing 2.0 Modeling and Metadata Strategies for Next Generation Architectures.* **Inmon, Bill. 2010.** s.l. : Technology, LLC, 2010.

23. *Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models.* **Breslin, Mari. 2004.** 2004, Business Intelligence Journal.

24. **de Armas Ramírez, Nerelys y Valle Lima, Alberto. 2001.** Resultados científicos en la investigación educativa. s.l. : Editorial Pueblo y Educación, 2001.

25. **ENISA. 2016.** ENISA Threat Landscape 2015. [En línea] 02 de 2016. www.enisa.europa.eu.

26. **Gilson, N, y otros. 2009.** *Development of a Framewordk for Workplace Intervention Using the Delphi Technique.* 2009. págs. 6,520-528.

27. **González Brito, Henry Raúl. 2016.** https://henryraul.wordpress.com. *https://henryraul.wordpress.com.* [En línea] 2016.

28. **González Hernández, Yanisbel. 2013.** *Metodología de desarrollo para proyectos de almacenes de datos.* La Habana : s.n., 2013.

29. **Gosaina, Anjana y Arorab, Amar . 2015.** Security Issues In Data Warehouse: A Systematic Review. [En línea] 2015. www.sciencedirect.com.

30. **Gravitar. 2016.** Gravitar. [En línea] 2016. [Citado el: 13 de Junio de 2016.] http://gravitar.biz/pentaho/.

31. **GSI - Facultad de Ingeniería. 2016.** *Seguridad Informática Identificación, Autenticación, Autorización.* 2016.

32. **help.pentaho.com.** https://help.pentaho.com. [En línea] [Citado el: 16 de Junio de 2016.]

33. *Importancia de la utilización de un Data Warehouse (DW) en las empresas.* **Rizo Rizo, Emma R, y otros. 2006.** 2006.

34. **Inmon, W. H., Imhoff, Claudia y Sousa, Ryan. 2001.** Corporate Information Factory. 2001.

35. **Inmon, William, Strauss, Derek y Neushloss, Genia. 2008.** *DW2.0 The Architecture for the Next Generation of Data Warehousing.* 2008.

36. **Kimball, R y Caserta, J. 2004.** *The Data Wharehouse ETL Toolkit.* 1. s.l. : Wiley Publishing, 2004.

37. **Kimball, Ralph y Caserta, Joe. 2004.** *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* 1ed. s.l. : 10475 Crosspoint Boulevard Indianapolis: Wiley Publishing, 2004.

38. **Kimberly, Madia. 2013.** Data Security Strategies to Keep the Bad Guys at Bay and the Good Guys Honest. *IBM Big Data & Analytics Hub.* [En línea] Junio de 2013.

39. *Método multicriterio multiexperto para evaluar el impacto de las TICs en la formación del ingeniero en Ciencias Informáticas.* **Jiménez Hernández, Regla C y Mar Cornelio, Omar. 2016.** La Habna : s.n., 2016.

40. *Metodología para la Extracción del Conocimiento Empresarial a partir de los Datos.* **Matos, Guillermo, Chalmeta, Ricardo y Coltell, Oscar. 2006.** 2, 2006, Vol. 17.

41. *Modeling security-relevant data semantics.* **Gary W, Smith. 1991.** 1991, IEEE Transactions on Software Engineering.

42. *Modelo de gestión del conocimiento de la investigación para Colombia y Cuba.* **Ortiz Bojacá, José Joaquín y Francisco, Borrás Atiénzar. 2012.** México, D.F. : s.n., 2012. XVII Congreso internacional de contaduría, administración e informática.

43. **Montesino Perurena, Raydel. 2013.** *Modelo para la gestión automatizada e integrada de controles de seguridad informática.* La Habana : s.n., 2013.

44. **Pentaho BI. 2011.** Pentaho explore endless possibilities. [En línea] 2011. [Citado el: 13 de Junio de 2016.] http://bievolutivo.com/es/home/pentaho.

45. **Pressman, Roger S. 2007.** *Ingeniería de Software.Un enfoque práctico.* 6ta. 2007. ISBN:8448132149.

46. **Ralph, Kimball y Ross, Margy. 2010.** The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence. Indianapolis, Indiana : Wiley Publishing, Inc, 2010.

47. *RBAC Extension Model for ERP Systems in Multi-domain Environments.* **Gomez Baryolo, Oiner , Estrada Senti, Vivian y Lazo Cortes, Manuel S. 2012.** s.l. : IEEE Latin America Transactions, 2012.

48. *Security and the Data Warehouse.* **Browder, Kristy y Lumpkin, George. 2005.** Abril de 2005.

49. *Security of Data Warehousing Server.* **Palletvuori, Kimmo. 2010.** 2010. Seminar on Network Security.

50. *Seguridad Informática Conceptos generales.* **Alonso Romero, Luis. 2002.** Universidad de Salamanca. España : s.n., 2002. Ponencia: III Jornadas Sobre Derecho e Informática: Protección de Datos y Seguridad en Internet.

51. **Siegel, S. 1974.** *Estadística no paramétrica : aplicada a las ciencias de la conducta.* s.l. : Trillas, 1974. págs. 262-273.

52. *Sistema informático para calcular el Índice de Control Interno en instituciones cubanas.* **Bron Fonseca, Barbara y Jiménez Hernández, Regla . 2016.** 2016.

53. *The SELinux Notebook - The Foundations.* **The GNU Free Documentation License. 2007.** 2007.

54. **The server labs. The IT architects. 2016.** The server labs. The IT architects. [En línea] 2016. http://www.theserverlabs.com/folletos/Folleto%20SOA.

55. **Trujillo, Juan Carlos, Mazón, José Norberto y Pardillo, Jesús. 2013.** *Diseño y explotación de Almacenes de datos.* s.l. : ECU, 2013. 9788499485461.

56. *Understanding LDAP Design and Implementation.* **IBM. 2004.** s.l. : IBM, 2004.

57. **Unicon. 2016.** Unicon. [En línea] 2016. [Citado el: 14 de Junio de 2016.] https://www.unicon.net/opensource/cas.

58. *View Security as the Basic for DW Security.* **Rosenthal , Arnon y Sciore, Edward. 2000.** 2000.

59. **www.cvedetails.com. 2016.** www.cvedetails.com. *www.cvedetails.com.* [En línea] 28 de Octubre de 2016.

GLOSARIO

Amenaza: evento que puede afectar los componentes de una aplicación informática a través de la explotación de una vulnerabilidad.

Arquitectura de seguridad: descripción detallada de todos los aspectos relativos a la seguridad de un sistema (o red), incluye un grupo de principios para guiar el diseño de la seguridad. La arquitectura describe cómo todos estos aspectos se integran para satisfacer los requerimientos de seguridad.

Ataque: una acción que se sirve de una o varias vulnerabilidades para materializar una amenaza.

Auditoría: se refiere a la capacidad de mantener un registro de las transacciones realizadas en una aplicación. La auditoría permite saber quién hizo qué, cuando lo hizo, y quién le proporcionó los permisos necesarios a ese usuario.

Autenticación: capacidad de validar la identidad de un usuario. Típicamente se realiza por medio de nombres de usuario y contraseña.

Autorización: es la definición de qué es lo que un usuario específico puede hacer dentro de una aplicación, es decir a qué información y operaciones tiene acceso.

Confidencialidad: ausencia de la divulgación no autorizada de la información.

Control: medios para manejar el riesgo, incluyendo políticas procedimientos, lineamientos, prácticas, estructuras organizacionales o medios técnicos que son aplicados para eliminar o reducir una vulnerabilidad.

Disponibilidad: correcta disposición de los servicios del sistema.

Incidente: evento no deseado que podría dañar o reducir el valor de los activos.

Integridad: ausencia de alteraciones del sistema.

Mecanismos de seguridad: un mecanismo de seguridad (también llamado herramienta de seguridad o control) es una técnica que se utiliza para implementar un servicio, es un mecanismo que está diseñado para detectar, prevenir o recobrarse de un ataque de seguridad.

Principios de seguridad: son las bases que deben seguirse para alcanzar niveles adecuados de seguridad.

Riesgo: combinación de la probabilidad de que se produzca un evento y su ocurrencia.

Vulnerabilidad: error presente en el software que puede ser usado por un atacante para acceder a una aplicación informática o red de datos, violando las políticas de seguridad informática establecidas.

ANEXO 1. Formatos de ejemplos

Tabla 10: Formato para la especificación de requisitos. Elaboración propia.

Requisitos Subsistema Fuente de datos	Requisitos Subsistema Procesos ETL	Requisitos Subsistema Almacén de Datos	Requisitos Subsistema Visualización de Datos
Se listan (R1...Rn) y detallan cada uno de los requisitos que deberán implementarse en el subsistema Fuente de datos.	Se listan (R1...Rn) y detallan cada uno de los requisitos que deberán implementarse en el subsistema Procesos ETL.	Se listan (R1...Rn) y detallan cada uno de los requisitos que deberán implementarse en el subsistema Almacén de Datos.	Se listan (R1...Rn) y detallan cada uno de los requisitos que deberán implementarse en el subsistema Visualización de Datos.
Ejemplo			
R1. Garantizar el acceso a la información de los sistemas fuente solo por las personas autorizadas.	**R2.** Garantizar la integridad de la información almacenada en los sistemas fuentes debido al nivel de sensibilidad de los mismos.	**R3.** El acceso para el mantenimiento del almacén de datos deberá realizarlo un usuario con permisos de administración en un período semanal emitiendo un informe del mismo.	**R4.** Existirán diferentes tipos de usuarios según las acciones que puedan realizar. Los permisos de los usuarios estarán en correspondencia con los niveles de acceso a la información clasificada almacenada en los libros de trabajos y reportes.

Tabla 11: Formato para la identificación de los datos. Elaboración propia.

Fuente de Datos	Medio de Almacenamiento	Volumen de Información	Metadatos de la fuente	Datos en "staging area"
Listado de las fuente de datos (utilizando el índice para la enumeración: **FD1…Rn**).	Se especifican los medios de almacenamiento, por ejemplo: base de datos privadas, bases de datos públicas, ficheros de texto plano y formato, entre otros.	Se especifica el volumen de la información: capacidad de almacenamiento que emplea y necesita.	Se especifican los metadatos que se utilizarán para trabajar con la fuente de datos especificada.	Se especifican los datos a almacenar en el área de almacenamiento intermedio "*staging area*", asociados a la fuente de datos especificada.
Ejemplo				
FD1. Datos de detenciones. **FD2.** Datos de las personas fallecidas. **FD3.** Datos asociados a expedientes delictivos. **FD4.** Datos relacionados	Todas las fuentes de datos se encuentran en fichero de texto plano con formato **.xml**	La fuente de datos **FD1** tiene un volumen de información de 100 MB, se necesita esta misma capacidad de almacenamiento. (Asimismo se especifica para el	Para la fuente de datos **FD1** se utilizará los metadatos: delitos, estado de detención y naturaleza del delito. Para la fuente de datos **FD3**	Se almacenarán en el área de almacenamiento intermedio información asociada a los datos de la fuente **FD4**.

con las personas solicitadas.		resto de las fuentes de datos).	se utilizará el metadato: estado del expediente delictivo.	

Tabla 12: Formato para la clasificación de los datos. Elaboración propia.

Conceptos de Información o Datos identificados	Clasificación
Se especifican los conceptos de información o datos identificados (utilizando el índice para la enumeración: **C1...Cn**) para el desarrollo del almacén de datos.	Se clasifica cada concepto de información o dato a almacenar en el AD en: públicos, confidenciales o secretos según su nivel de sensibilidad. Lo cual se utilizará para establecer las medidas de protección de datos.
Ejemplo	
C1. Detenciones. **C2.** Personas fallecidas. **C3.** Expediente delictivo. **C4.** Solicitados.	Se clasifican en confidenciales el concepto de información **C2**. Se clasifican en secretos los conceptos de información: **C1**, **C3** y **C4** porque su divulgación ofrece peligro para la seguridad ciudadana.

Tabla 13: Formato para la identificación de vulnerabilidades. Elaboración propia.

Vulnerabilidades Subsistema Fuente de datos	Vulnerabilidades Subsistema Procesos ETL	Vulnerabilidades Subsistema Almacén de Datos	Vulnerabilidades Subsistema Visualización de Datos
Se especifican las vulnerabilidades asociadas al subsistema Fuente de datos (utilizando el índice para la enumeración: **VFD1…VFDn**).	Se especifican las vulnerabilidades asociadas al subsistema Procesos ETL (utilizando el índice para la enumeración**:** **VETL1…VETLn**).	Se especifican las vulnerabilidades asociadas al subsistema Almacén de Datos (utilizando el índice para la enumeración: **VAD1…VADn**).	Se especifican las vulnerabilidades asociadas al subsistema Visualización de Datos (utilizando el índice para la enumeración: **VVD1…VVDn**).
Ejemplo			
VFD1. Las fuentes de datos se encuentran en ficheros de texto plano almacenados en disco duro sin protección.	**VETL1.** Comunicación no segura entre los subsistemas Procesos ETL y Fuente de datos.	**VAD1.** Fallos en el sistema operativo donde se encuentra el almacén de datos.	**VVD1.** El sistema de visualización no implementa políticas de acceso a los libros de trabajo y reportes que incluye la capa de visualización de datos.

Tabla 14: Formato para las medidas de protección de datos. Elaboración propia.

Medidas de protección Subsistema Fuente de datos	Medidas de protección Subsistema Procesos ETL	Medidas de protección Subsistema Almacén de Datos	Medidas de protección Subsistema Visualización de Datos
Se especifican las medidas de protección de datos (utilizando el índice para la enumeración: **M1…Mn**) para cada vulnerabilidad especificada en el subsistema Fuente de datos, ejemplo medida 1 (M1) para mitigar la vulnerabilidad (VFD1): **M1-->VFD1**: Descripción de la medida. Pueden especificarse varias medidas de seguridad para cada vulnerabilidad.	Se especifican las medidas de protección de datos (utilizando el índice para la enumeración: **M1…Mn**) para cada vulnerabilidad especificada en el subsistema Procesos ETL, ejemplo medida 2 (M2) para mitigar la vulnerabilidad (VETL1): **M1--> VETL1**: Descripción de la medida. Pueden especificarse varias medidas de seguridad para cada vulnerabilidad.	Se especifican las medidas de protección de datos (utilizando el índice para la enumeración: **M1…Mn**) para cada vulnerabilidad especificada en el subsistema Almacén de Datos, ejemplo medida 3 (M3) para mitigar la vulnerabilidad (VAD1): **M1--> VAD1**: Descripción de la medida. Pueden especificarse varias medidas de seguridad para cada vulnerabilidad.	Se especifican las medidas de protección de datos (utilizando el índice para la enumeración: **M1…Mn**) para cada vulnerabilidad especificada en el subsistema Visualización de Datos, ejemplo medida 4 (M4) para mitigar la vulnerabilidad (VVD1): **M1--> VVD1**: Descripción de la medida. Pueden especificarse varias medidas de seguridad para cada vulnerabilidad.
Ejemplo			
M1-->VFD1. Se protegerá el acceso a la fuente de datos mediante mecanismos de seguridad desde el sistema operativo, con el módulo de seguridad *SELinux* en el servidor “Fuente de Datos”.	**M1--> VETL1.** Se garantizará la integridad de la información de la fuente de datos mediante el cálculo de valores hash de cada fichero de texto plano de la fuente de datos.	**M1--> VAD1.** Se realizarán salvas de la base de datos que representa el almacén de datos con periodicidad semanal.	**M1--> VVD1.** Se implementarán políticas de usuarios y permisos para el acceso a los libros de trabajo y reportes del subsistema de visualización de datos.

Tabla 15: Formato para la evaluación de las medidas de seguridad. Elaboración propia.

Nombre del Caso de Prueba: Se especifica el nombre del Caso de prueba, por ejemplo: caso de prueba 1 que ejecute la vulnerabilidad 1 relacionada con el subsistema Fuente de datos: CP1-->VFD1 Nombre del Caso de Prueba.			
Escenario	Flujo central	Variables de entradas	Respuesta
Se especifica el nombre de cada escenario del Caso de prueba.	Se describe el flujo de actividades a llevar a cabo para ejecutar el escenario del Caso de prueba.	Se especifican las variables o elementos de entrada y sus respectivos valores. Pueden establecerse columnas para cada variable o elemento y establecer sus respectivos valores.	Se describe la respuesta luego de ejecutar el escenario del Caso de prueba.
Ejemplo			
CP1-->VFD1. Acceder a la fuente de datos.			

SC 1. Acceder a ficheros del sistema fuente de datos con permiso.	Este es el escenario ideal donde todos los datos de las variables son correctos y consecuentemente es exitosa la ejecución de la prueba. Al introducir un nombre de usuario y contraseña con permisos válidos, dicho usuario debe acceder satisfactoriamente a la dirección del fichero.	Usuario: administrador Contraseña: Process1ETL*D **Directorio completo del fichero:** /Datos/Datos_Detenciones/listado_detenciones.xml	El usuario autenticado con permisos válidos accede satisfactoriamente a la dirección del fichero.
SC 1. Acceder a ficheros del sistema fuente de datos sin permisos.	Este es el escenario donde los datos de las variables no son correctos y consecuentemente no se ejecuta satisfactoriamente el caso de prueba.	Usuario: administrador Contraseña: Process1 **Directorio completo del fichero:** /Datos/Datos_Detenciones/listado_detenciones.xml	El usuario autenticado no accede a la dirección del fichero.

ANEXO 2. Diagramas de componentes

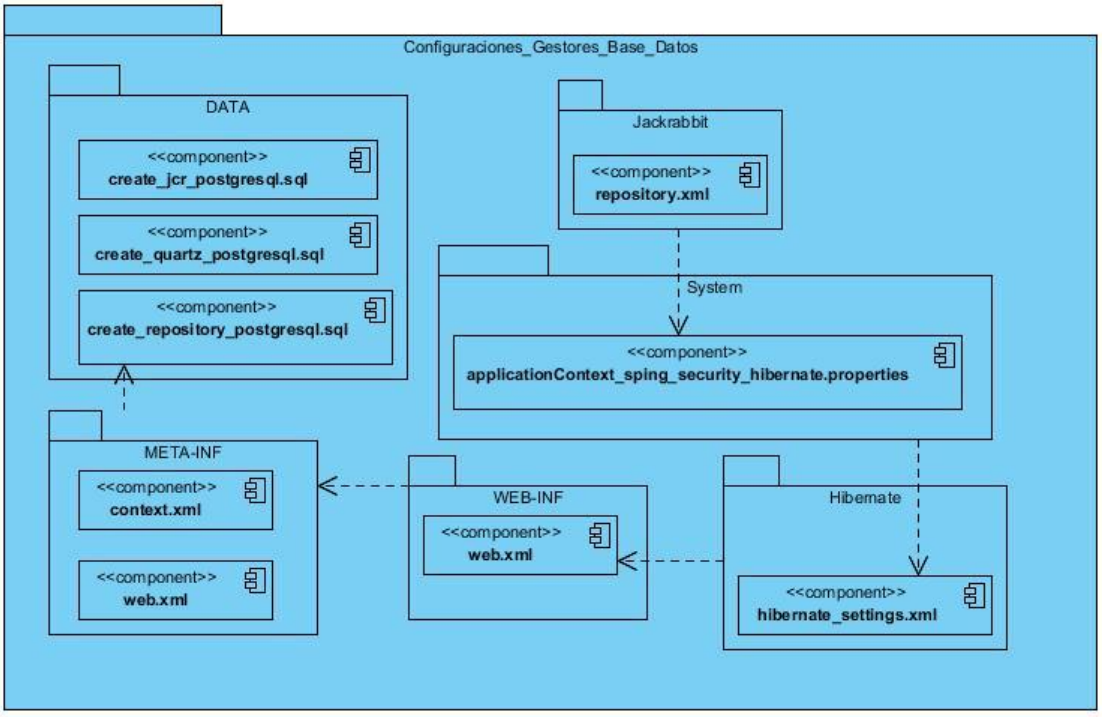


Figura 42: Paquete: Configuraciones_Gestores_Base_Datos. Elaboración propia.

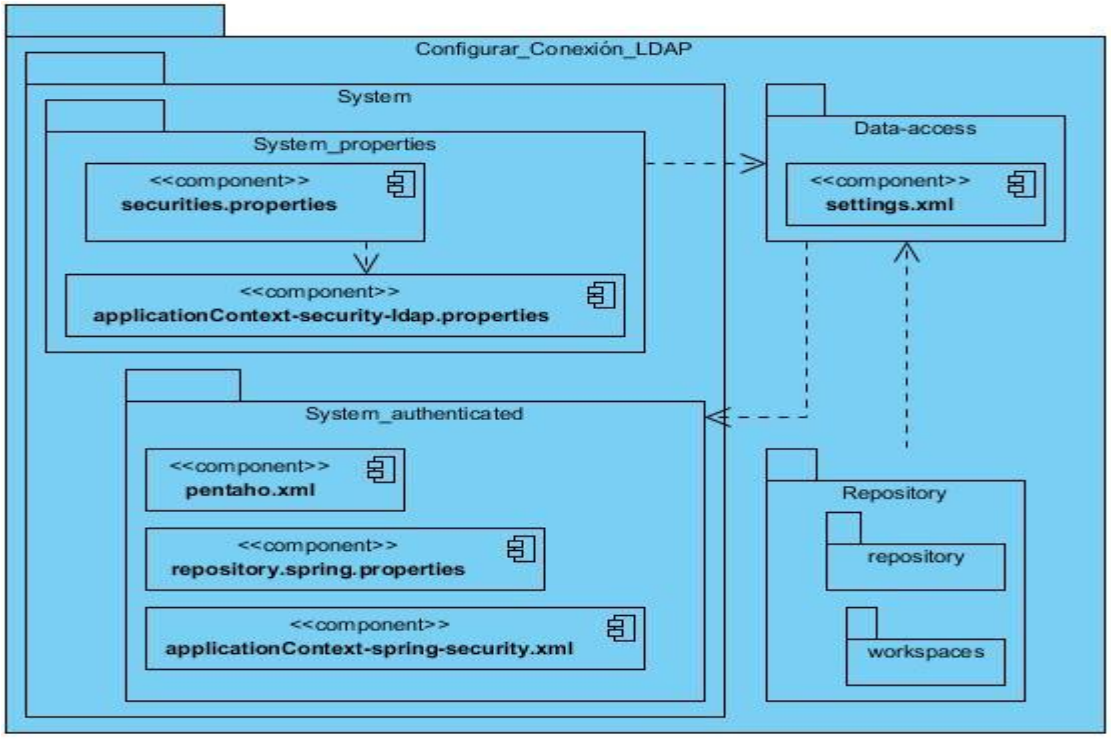


Figura 43: Paquete: Configurar Conexión LDAP. Elaboración propia.

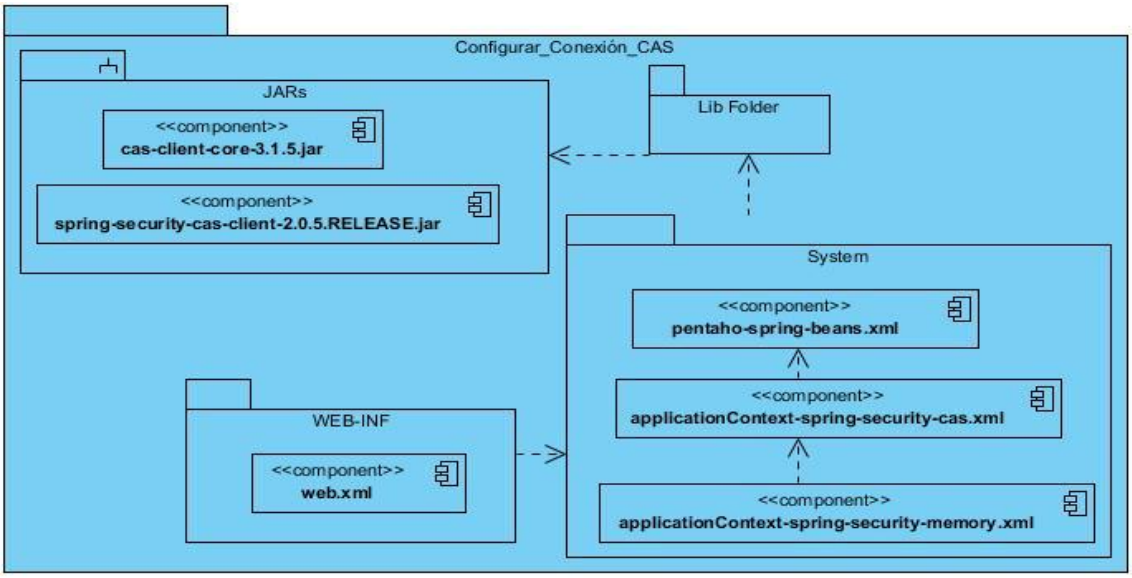


Figura 44: Paquete: Configurar Conexión CAS. Elaboración propia.

ANEXO 3. Encuesta. Nivel de experiencia de los expertos

Nombre (s) y Apellidos:

Labor que realiza:

Calificación profesional: Ingeniero___ Licenciado __ Máster___ Doctor___

Categoría docente: Prof. Instructor___ Prof. Asistente___ Prof. Auxiliar ___ Prof. Titular___ Prof. Adjunto___

1. Seleccione en una escala del 1 al 10 el valor que corresponda con el grado de conocimientos que usted posee acerca del tema de investigación que se desarrolla (seguridad de aplicaciones), considerando 1 como no tener ningún conocimiento y 10 el de pleno conocimiento de la problemática tratada.

1	2	3	4	5	6	7	8	9	10

2. Valore el grado de influencia que cada una de las fuentes que se le presenta a continuación ha tenido en su conocimiento y criterios sobre el tema que se investiga.

Fuentes de argumentación	Grado de influencia de cada una de las fuentes:		
	Alto	**Medio**	**Bajo**
Análisis teóricos realizados			
Experiencia obtenida.			
Conocimiento de trabajos de autores nacionales.			
Conocimiento de trabajos de autores extranjeros.			
Conocimiento del estado del problema internacionalmente.			
Su intuición.			

ANEXO 4. Encuesta. Medidores de los resultados de la investigación

Estimado experto:

Usted ha sido seleccionado por su calificación científica-técnica, sus años de experiencia y sus resultados profesionales, como experto para seleccionar de acuerdo a su opinión, los posibles criterios de evaluación para esta investigación, partiendo de la siguiente propuesta. La evaluación del criterio, significa su valoración respecto a ese parámetro como medidor de la investigación, el mismo puede ser: Muy bueno, Bueno, Regular, Pobre y Malo. Exprese su opinión en un valor numérico de 1 a 5, siendo 5 Muy bueno y 1 Malo.

Si usted considera que el criterio está mal categorizado, haga una nueva propuesta. De igual manera, si considera que debe incluir otros criterios.

Categoría de los criterios	Propuesta de peso para la categoría	Criterio	Evaluación del criterio **(de 1 a 5)**	Otra propuesta de categorización	Observación
Científica		Calidad de la investigación			
		Novedad científica			
		Coherencia entre el fundamento teórico y la propuesta			
Necesidad		Necesidad de la estrategia de seguridad			
		Satisfacción de las necesidades productivas			
Aplicabilidad		Facilidad de comprensión de la estrategia de seguridad propuesta			
		Adaptabilidad a diferentes soluciones de almacenes de datos del centro DATEC de la UCI			
		Facilidad de uso			
		Impacto			
		Cumplimiento de principios de seguridad			
Económico		Contribución a disminuir los costos de desarrollo			
		Garantía de seguridad			

Calidad técnica		Orden y estructura correcta de la estrategia de seguridad propuesta			
Nuevas categorías y/o criterios					

ANEXO 5. Encuesta. Valoración de los resultados de la investigación

Estimado experto:

Usted ha sido seleccionado por su calificación científica-técnica, sus años de experiencia y sus resultados profesionales, como experto para evaluar los resultados teóricos de la investigación "Estrategia de seguridad para soluciones de Almacenes de datos". Le agradecemos sus criterios acerca de las ventajas, deficiencias e insuficiencias que presenta la estrategia de seguridad en su concepción teórica y su futura aplicación. Sus criterios y opiniones se analizan de forma anónima. Se agradece su valiosa colaboración.

Emita su opinión acerca de la importancia que le atribuye a cada categoría con relación al resto. La sumatoria de estos valores debe ser igual a 100.

Categoría de los criterios	Criterios	Propuesta de peso (Sumatoria igual a 100)
Científica	Calidad de la investigación	
	Novedad científica	
	Coherencia entre el fundamento teórico y la propuesta	
Necesidad	Necesidad de la estrategia de seguridad	
	Satisfacción de las necesidades productivas	
Aplicabilidad	Facilidad de comprensión de la estrategia de seguridad propuesta	
	Adaptabilidad a diferentes soluciones de almacenes de datos del centro DATEC de la UCI	
	Facilidad de uso	
	Impacto	
	Cumplimiento de principios de seguridad	
Económico	Contribución a disminuir los costos de desarrollo	
	Garantía de seguridad	
Calidad técnica	Orden y estructura correcta de la estrategia de seguridad propuesta	

ANEXO 6. Resultados de la evaluación. Pesos otorgados

Tabla 16: Pesos otorgados por los expertos. Elaboración propia.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
E1	5	4	5	5	5	5	5	5	5	5	4	5	5
E2	4	5	5	5	5	5	5	5	5	4	3	4	5
E3	5	4	5	5	5	5	5	5	4	5	5	4	5
E4	5	5	5	5	4	5	4	5	5	4	5	4	5
E5	4	4	5	4	4	5	5	5	4	4	4	4	5
E6	5	4	5	5	5	4	5	4	5	5	3	4	4
E7	5	4	5	5	5	5	5	5	4	5	4	5	5
E8	5	5	5	5	5	5	5	5	5	4	5	5	5
E9	5	4	5	5	5	5	5	5	5	5	3	5	5
E10	5	4	5	5	5	4	5	4	4	5	4	4	5
∑E	48	43	50	49	48	48	49	48	46	46	40	44	49
EP	4.8	4.3	5	4.9	4.8	4.8	4.9	4.8	4.6	4.6	4	4.4	4.9