Auditoría a bodega de datos del Centro de Datos e Indicadores Ambientales de Caldas (CDIAC) ubicado en la Universidad Nacional de Colombia

Valentina Olaya

Trabajo de grado para optar por el título de:
Administrador de Sistemas Informáticos

Profesor asesor:
Néstor Darío Duque Méndez

Universidad Nacional de Colombia
Departamento de Informática y Computación

Manizales
Febrero 2020

Introducción

El Centro de Datos e Indicadores Ambientales (CDIAC) es una plataforma web que permite visualizar la información ambiental del departamento de Caldas. Este proyecto crea la bodega de datos ambientales que soporta la plataforma CDIAC, permitiendo la disponibilidad de indicadores ambientales para su consulta a la comunidad [1].

Actualmente dicha bodega cuenta con más de 50 millones de registros ambientales, tomados desde hace aproximadamente 20 años, convirtiendo así a esta bodega no solo importante para el proyecto CDIAC si no también para todo el departamento Caldense, de ahí la importancia de contar con una bodega protegida y segura.

En este trabajo de grado se pretende proponer una metodología para realizar una auditoría basada en riesgos a dicha bodega, con el fin de evaluar los riesgos en los dominios de esta, y así poder diseñar controles que minimicen la probabilidad de ocurrencia de los riesgos encontrados.
Al final de este trabajo se encontrará un informe con sus respectivas recomendaciones y conclusiones.

Objetivo general

Proponer una metodología para realizar auditoría basada en riesgos a la bodega de datos del Centro de Datos e Indicadores Ambientales (CDIAC).

Objetivos específicos

1. Definir los dominios relacionados con los componentes de la Bodega de Datos.
2. Identificar posibles riesgos de la Bodega de Datos para cada dominio.
3. Definir mecanismos de control que puedan implementarse para los riesgos hallados.
4. Realizar evaluación de controles en la bodega del CDIAC implementando Técnicas de Auditoría con Ayuda de Computadora (TAAC).
5. Realizar informe final para la Bodega de Datos de CDIAC.

CAPÍTULO 1

Marco Teórico

Bodega de datos

Una bodega de datos (data warehouse) es un almacén de datos integrados, no volátil y orientados a una materia, estos datos están disponibles para consultas y su objetivo es soportar la toma de decisiones oportunas y fundamentadas en la organización a la que se aplica [2].

Los datos de estas bodegas se obtienen de diferentes fuentes como bases de datos, archivos planos u otras plataformas web, y deben ser sometidos a un proceso de extracción, transformación y carga (ETL) para garantizar la veracidad, integridad y centralización. El ETL permite resolver diferentes tipos de problemas como pueden ser unidades de medida, codificaciones, fuentes de datos múltiples, entre otros. Estos datos suelen cubrir periodos históricos y ser de gran volumen [3].

Una bodega de datos se dice que es orientada a un tema o materia porque la información es en base a los intereses que tenga la organización para la que fue creada, permitiendo un mejor manejo y control de la información y es integrada cuando los datos son producidos por diferentes aplicaciones, ya sean externas o internas a la compañía [3].

Dominios

COBIT define los dominios como la agrupación de procesos que corresponden a un conjunto de actividades unidas con delimitación o cortes de control y objetivos de control, con el objetivo de lograr resultados medibles y de interés para la compañía.

Desde COBIT 4.0 se definen las actividades clasificadas por dominios, para facilitar la respectiva revisión de cada proceso, en un principio se definieron cuatro dominios, pero en la última versión se definen cinco dominios, los cuales son:

1. Evaluar, orientar y supervisar
2. Alinear, planificar y organizar
3. Construir, adquirir e implementar
4. Entregar, dar servicio y soporte
5. Supervisar, evaluar y valorar

Estos dominios cuentan con un total de 37 procesos [4].

Para la metodología propuesta en este trabajo de grado, se decide basarse en la estructura adoptada por COBIT y definir los propios dominios de una bodega de datos, los cuales se mencionan más adelante.

Riesgos

Según la guía ISO 73 el riesgo se define como la combinación de la probabilidad de un suceso y de su consecuencia, este término debe ser usado solo en caso de que exista, al menos, una posibilidad de consecuencia negativa, en caso de obtener una consecuencia positiva no se considera riesgo. Puede ocurrir que el riesgo surja de la posibilidad de desviación con respecto al resultado o suceso previsto [5].

Riesgo Inherente

El riesgo inherente es el resultado del producto entre los valores probabilidad de ocurrencia e impacto del riesgo, el valor resultante se utiliza para clasificar los riesgos en un nivel de importancia una evaluación del riesgo.

Una evaluación de riesgo consiste en comparar el riesgo calculado con ciertos criterios de riesgo, para así poder determinar la importancia de este [5]. Para este caso los criterios definidos son la probabilidad de ocurrencia y el impacto del riesgo.

Controles

Para la guía ISO 73 los controles son definidos como acciones que ponen en aplicación las decisiones de la gestión del riesgo, dichos controles pueden incluir supervisión, reevaluación y conformidad con las decisiones tomadas en la compañía [5].

Auditoría basada en riesgos

La auditoría basada en riesgos consiste en ampliar la perspectiva de la auditoría para incluir cualquier técnica que permita identificar un riesgo, inclusive técnicas administrativas. Esta práctica le proporciona al auditor la posibilidad de de examinar si existe un excesivo control en los procesos y así poder identificar procedimientos que se encuentren obsoletos o sean ineficientes, en términos generales, la auditoría basada en riesgos trata de conocer y entender los factores que puedan tener efectos desfavorables en las operaciones del negocio [6].

Datos Ambientales

Los datos ambientales comprenden información sobre los elementos del medio ambiente, como lo son el aire, agua o clima. Con dichos datos se puede conocer el estado de cada factor ambiental y que contaminantes, vertientes o emisiones pueden afectar o alterar su estado natural.
CDIAC permite el acceso a datos ambientales del departamento de Caldas, permitiendo que la población en general pueda estar al tanto del estado ambiental del departamento, y genera indicadores para facilitar el entendimiento de las variables ambientales tomadas [1].

CAPÍTULO 2

Trabajos relacionados

A continuación, se presentan algunos de los trabajos encontrados relacionados con el desarrollo de una metodología que ayude a la auditoría de una bodega de datos, estos trabajos relacionados amplían la perspectiva de la importancia que tiene el aseguramiento de las bodegas en la actualidad.

En [7] el proceso de auditoría plantea verificar el sistema de control ya establecido con el fin de proporcionar a la gerencia la seguridad de que están cumpliendo con los objetivos de control, de igual manera corroborar los riesgos causados por alguna debilidad en un control implementado y realizar un asesoramiento en acciones correctivas.
La metodología propuesta en este trabajo se basa en la evaluación de riesgos y a partir de riesgos inherentes que amenacen el proyecto, se establecen los objetivos de control que minimicen las amenazas previamente identificadas, con el fin de tener información que satisfaga los requisitos del negocio.

El documento [8] presenta un sistema experto que respalda al auditor de sistemas en la evaluación de riesgos y la elección de controles que protegen razonablemente a la organización. La base de conocimiento representa los hechos dados y las acciones del experto en auditoría de sistemas bajo la metodología de análisis de riesgos. Este sistema se implementa con un software gratuito en un entorno web, como objetivo esta propuesta tiende a trasladar el conocimiento y la experiencia a diferentes espacios donde el auditor no puede generar permanentemente un ahorro de tiempo y recursos. Este desarrollo implica tres etapas: La definición del programa de auditoría, El diseño del sistema experto y la Implementación de la aplicación.

CAPÍTULO 3

Metodología propuesta

Para realizar la auditoría a la bodega de datos se propone una auditoría basada en riesgos de los dominios de la bodega, por lo cual, en primer lugar, se deben identificar y definir dichos dominios.

Después se realiza la identificación de riesgos en cada uno de los dominios identificados apoyados de una revisión bibliográfica, también se debe realizar la evaluación de cada riesgo, con el fin de poder diseñar los respectivos controles para estos riesgos, como trabajo final se espera aplicar la metodología a la bodega de datos ambientales de CDIAC y poder realizar un informe final en el cual se evaluarán los riesgos existentes para esta bodega y verificar la existencia o funcionamiento de controles y si es necesario realizar recomendaciones necesarias.

A continuación, la figura 1 muestra el proceso de la metodología propuesta en este trabajo de grado.

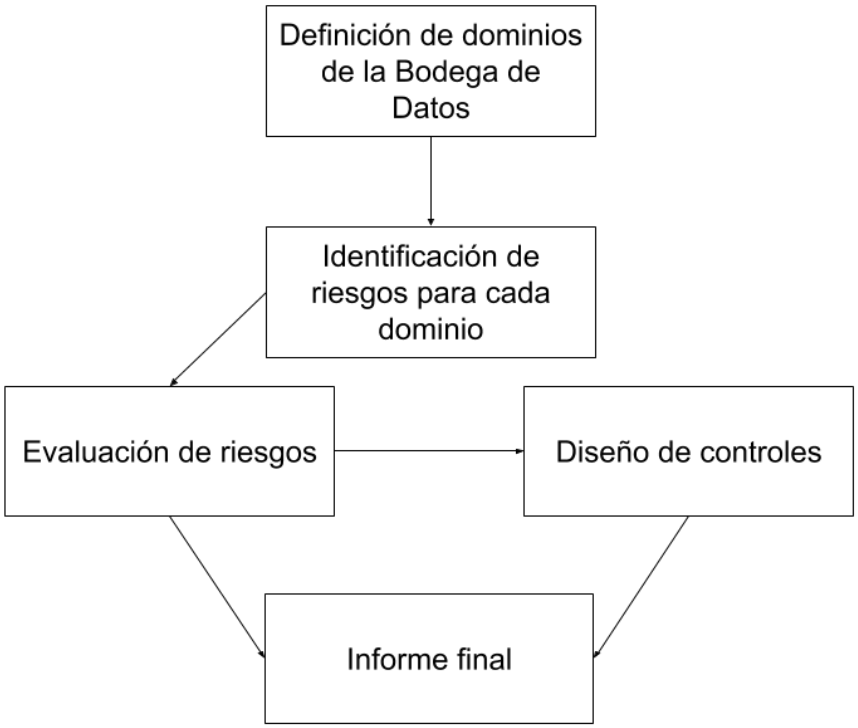

Fig. 1. Estructura de la metodología propuesta
Fuente: Elaboración propia

Definición de dominios de la bodega de datos a auditar

Los dominios que se plantean auditar fueron seleccionados en base a la restricción de tiempo que se tiene para realizar el trabajo de grado, ya que existen dominios cuya auditoría requiere más investigación y tiempo, debido a sus extensos o complejos procesos, por lo tanto, los mencionados a continuación no son los únicos dominios que se pueden definir para una bodega de datos.

Diseño e implementación

En esta fase se debe partir de los requerimientos funcionales de información, que generen una ventaja competitiva para la empresa y faciliten la toma de decisiones por parte de la administración, y los requisitos de las fuentes de datos operacionales, a partir de esto se debe obtener un modelo lógico basado en una tecnología de base de datos específica que cumplan con la función de guiar la implementación [3].

La complejidad de un almacén de datos aumenta geométricamente con la cantidad de datos que se ingresan en él, por lo cual el diseño de esta misma debe ir adaptándose a nuevos tamaños.

En la actualidad no se define un mecanismo específico que permita estructurar de una manera sistemática el diseño y desarrollo de una bodega de datos, convirtiendo esto en una tarea compleja y artesanal, siendo el diseño de la bodega, fundamental en la calidad del análisis de los datos que esta contiene [3]. Por lo tanto, se debe prestar atención a posibles problemas desde esta fase inicial.

Extracción, Transformación y Carga (ETL)

Los procesos de ETL permiten realizar el traslado desde el origen de los datos hasta la bodega, se encargan de extraer los datos de sus diferentes fuentes, integrarlos y finalmente filtrarlos y depurarlos.

La fase ETL es considerada como una etapa crucial en el almacenamiento de datos y donde reside la responsabilidad de obtener datos de calidad, este proceso aparte de permitir la validación de la calidad de los datos, también ayuda a auditar o rastrear problemas [9].

Otro componente fundamental en este proceso son los extractores de datos, los cuales se encargan del copiado y distribución de los datos de acuerdo con el diseño planteado por el ETL, en este componente se determinan los datos a copiar, desde donde y hacia dónde van, así como los periodos para las actualizaciones. Aquí se permite que los datos externos sean adecuados y limpiados antes de ser sumados a la bodega de datos.

Una de las causas de los fallos en la construcción de una bodega de datos es la falta de comprensión de los datos de origen al momento de su extracción [9], esto convierte a las herramientas de extracción de datos en algo importante y fundamental a la hora de decidir diseñar una bodega y es algo que en ocasiones no se tiene en cuenta en la planificación, debido a esto, se resaltó la importancia de este proceso en este trabajo de grado.

Bodega de datos

Corresponde al repositorio de datos actual el cual debe estar organizado y orientado a intereses concretos, estos repositorios deben contener la información detallada y agregada de la bodega de datos.

Actualmente la bodega de datos del CDIAC se encuentra en el sistema de base de datos relacional PostgreSQL, un repositorio de datos de objetos de código abierto el cual usa el lenguaje SQL, escalando de manera segura las cargas de trabajo.
Este repositorio incorpora diferentes herramientas, como lo son: Data Definition Language (DDL), Triggers, Indexing Costraints entre otros [10].

El repositorio de datos es la base para mantener una bodega de datos sólida y segura por lo tanto debe existir un constante monitoreo al funcionamiento y rendimiento de esta herramienta.

Metadatos

Los metadatos llevan registros de los datos almacenados, integrados en la misma base de datos, describen el contenido de los objetos de la bodega de datos: las tablas, índices y el contenido de los datos. Los metadatos definen los formatos, significado y origen de los datos y facilitan el acceso y administración a los datos en la bodega.

En los metadatos se debe describir cambios únicos en los datos, lo que quiere decir que si se realiza alguna modificación, es necesario registrar dicho cambio. Se debe tener en cuenta que los metadatos se utilizan desde la fase de diseño hasta la fase de implementación, lo que convierte este proceso en una actividad importante para los procesos de pruebas [11].

Los metadatos son importantes a lo largo del ciclo de vida de la bodega, ya que toda la información con relevancia se registra en la base de metadatos, por lo cual, si hay algún problema en esta base de datos, también se tendrán con la operación de la bodega [11], incluso el proceso de ETL, ya que los metadatos se suelen utilizar para la recuperación del este proceso en caso de ser necesario.

Herramientas de análisis y minería

Las herramientas de análisis de datos parten de la información ya almacenada en la bodega, con el fin de entregar resultados desde diferentes puntos de vista y según las necesidades del negocio, permitiendo diversos niveles de detalle [2]. Mientras que la minería de datos es el proceso que se encarga de buscar patrones significativos en los datos, que expliquen eventos pasados, con el objetivo de usar dichos patrones para ayudar a predecir eventos futuros.

Existen dos tipos de minería, los cuales son:

- Descriptiva: Aquí la minería se encarga del análisis de Información dimensional.
- Predictiva: Mientras que la minería predictiva se encarga de generar modelos, lo que es realmente minería de datos donde se parte de un cúmulo de datos con el fin de descubrir relaciones ocultas y complejas a partir de diversas operaciones [2].

Usuarios

Un usuario puede referirse tanto a un sistema como a cualquier persona que realice actividades en los objetos de la bodega de datos. Se tienen diferentes tipos de usuarios como:

Usuarios administradores: estos usuarios cuentan con permisos especiales que permiten realizar cambios en la bodega de datos, estos usuarios suelen ser pocos y confiables para la compañía, ya que su poder sobre la bodega es alto lo que los convierte en un riesgo importante.

- Usuarios consultores: son usuarios con capacidad de consultar los datos almacenados, estos usuarios no tienen permitido realizar cambios en la bodega.
- Otros sistemas: los sistemas que interactúan con la bodega de datos se consideran también usuarios, ya que estos suelen extraer datos de la bodega y procesarlos, ya sea para realizar análisis sobre ellos o mostrarlos a los demás usuarios de forma diferente.

Identificación de Riesgos para cada dominio

Diseño e implementación

Para el dominio de diseño e implementación se definieron los riesgos presentados en la tabla 1.

	Riesgo	Descripción
1	Identificación incorrecta de hechos, dimensiones o relaciones de la bodega	La identificación incompleta o incorrecta de hechos o dimensiones, o las relaciones individuales contribuyen a los problemas de calidad de los datos.
2	Relaciones de datos inapropiadas entre tablas	La selección inadecuada de la granularidad de registros puede conducir a un diseño de esquema deficiente y, por lo tanto, afectar la calidad de los datos.

Tabla 1. Descripción de riesgos para el dominio de diseño e implementación
Fuente: Elaboración propia.

Extracción, Transformación y Carga (ETL)

Para el dominio de ETL se definen los riesgos presentados en la tabla 2

	Riesgo	Descripción
3	Pérdida de datos en el proceso ETL	La pérdida de datos durante el proceso ETL (registros rechazados) causa problemas de calidad, integridad o confianza de los datos.
4	Valores nulos no encontrados en proceso de ETL	Los valores nulos no controlados en el proceso ETL causan problemas de calidad de datos.
5	La imposibilidad de reiniciar el proceso ETL desde los puntos de control sin perder datos	Un proceso de ETL que no cuente con la opción de recuperar el proceso desde cualquiera de sus fases puede causar pérdida o baja calidad en los datos.
6	ETL no eficiente debido a un mal diseño	Un mal planteamiento y diseño de ETL no permite una migración exitosa en la bodega de datos.
7	Los datos no se cargan completos a la bodega	Los datos de origen son más que los mostrados en la bodega de datos.

8	Problemas de calidad de los datos en las fuentes de datos	Pueden ser, errores en entradas de datos o errores en actualización de datos por un sistema humano o informático [9].
9	Selección inadecuada de las fuentes de datos candidatas	La selección inadecuada de las fuentes de datos candidatas causa problemas de calidad de los datos (fuentes que no cumplen con las reglas comerciales) .
10	Cambios inesperados en los sistemas de origen	Los cambios inesperados en los sistemas de origen causan problemas de calidad de datos.
11	Uso de diferentes formatos de representación en fuentes de datos	Diferentes formatos para los mismos datos en las fuentes.
12	Presencia de registros duplicados	La presencia de registros duplicados de los mismos datos en múltiples fuentes causa problemas de calidad de datos.
13	Valores faltantes en las fuentes de datos	En las fuentes de datos pueden resultar fechas sin datos.
14	Datos mal escritos o en medidas erróneas	Valores en la bodega con errores de ortografía o medidas incorrectas.
15	Columnas adicionales	La bodega puede presentar columnas adicionales innecesarias.
16	Valores predeterminados	Valores predeterminados como variables utilizados para datos faltantes.
17	No especificar el carácter NULL en fuentes de datos	No especificar el carácter NULL correctamente en las fuentes de datos de archivos planos da como resultado datos incorrectos.
18	Valores atípicos	Valor distante de los valores normalmente presentados.
19	Extracción incorrecta de datos	La extracción incorrecta de datos a los campos requeridos causa problemas de calidad de datos.

20	No actualizar las fuentes de datos	No actualizar las fuentes de manera oportuna causa problemas de calidad de datos.
21	No actualizar réplicas de datos	Si no se actualizan todas las réplicas de datos, se producen problemas de calidad de datos.

22	Actualizaciones no documentadas	Cada que se realice una actualización en el sistema se espera que esta sea documentada.
23	Falta de actualización periódica	La falta de actualización periódica del almacenamiento de datos integrado causa la degradación de la calidad de los datos.
24	Malas conversiones del sistema	Las malas conversiones del sistema, la migración, la reingeniería o la consolidación contribuyen a los problemas de calidad de los datos.
25	Columnas que tienen valores de datos incorrectos	Columnas que tienen valores de datos incorrectos o formato de datos inconsistentes un ejemplo puede ser: el nombre de una persona se almacena en una tabla en el formato "Nombre Apellido" y en otra tabla en el formato "Apellido, Nombre".
26	Riesgo semántico	Se presenta cuando en los datos de origen y destino una columna tiene el mismo nombre, pero los datos se encuentran en medidas diferentes.
27	Corrupción de datos	Corrupción de datos, se refiere al cambio de formato o contenido original en la bodega de datos, este riesgo puede surgir en la migración de los datos, pueden contener anomalías, redundancias o algunos datos pueden estar duplicados.

Tabla 2. Descripción de riesgos para el dominio de ETL
Fuente: Elaboración propia.

Herramientas de análisis y minería de datos

A continuación, en la tabla 3 se presentan los riesgos identificados para las herramientas de análisis y minería de datos.

	Riesgo	**Descripción**
28	Mal uso de herramientas de análisis y minería de datos	El uso indebido de estas herramientas puede poner en duda la credibilidad de los datos.
29	Dificultad en la recopilación de los datos	Puede suceder que la recopilación de datos requiere mucho trabajo o incluso la inversión en tecnología sea de un precio elevado.

Tabla 3. Descripción de riesgos para el dominio de herramientas de análisis y minería
Fuente: Elaboración propia.

Usuarios

Para el dominio usuarios, se identificaron los riesgos presentados en la tabla 4.

	Riesgo	Descripción
30	No definir los perfiles para los usuarios	No tener perfiles definidos causa confusiones en los permisos que deben tener los diferentes usuarios.
31	Usuarios con permisos inadecuados	Los usuarios cuentan con permisos que permiten realizar acciones no necesarias para el desarrollo de su trabajo correspondiente.
32	Usuarios que abusen de los permisos asignados para el desarrollo de su trabajo	Pueden existir usuarios con permisos de administrador que quieran hacer uso de ellos para beneficio personal o de terceros.
33	Personal poco capacitado	Personal con poco conocimiento en las tecnologías que se usan o con el dominio de bodegas de datos y que cuentan con permisos para acceder a las herramientas.
34	Actualización de perfiles	La compañía debe mantener una actualización constante de los perfiles y sus respectivos permisos.

Tabla 4. Descripción de riesgos para el dominio de usuarios
Fuente: Elaboración propia.

Bodega de Datos

La tabla 5 muestra los riesgos identificados para este dominio.

	Riesgo	Descripción
35	Inyección SQL	La inyección SQL es un ataque que contiene código malicioso que se puede generar en la bodega de datos y que permite al atacante acceder y obtener o modificar los datos allí almacenados [12].
36	Truncar tablas	Truncar el área de almacenamiento de datos causa problemas de calidad de datos porque no podemos recuperar los datos para conciliarlos.
37	Espacio lógico	No hay espacio suficiente para almacenar la información de la bodega de datos.

38	Espacio Físico	Se debe contar con un espacio adicional en las tablas por si no hay suficiente capacidad de almacenamiento.

Tabla 5. Descripción de riesgos para el dominio de bodega de datos
Fuente: Elaboración propia.

Metadatos

En la tabla 6 se muestran los riesgos para los metadatos.

	Riesgo	**Descripción**
39	No tener metadatos	La falta de un repositorio de metadatos centralizado conduce a una mala calidad de los datos.
40	Valores predeterminados que no se cargan correctamente	Los valores predeterminados se definen y almacenan en la base de datos de metadatos y, por lo tanto, no debemos olvidar comprobar si estos valores se cargan correctamente.
41	Marca de tiempo no establecida para registros	En el proceso de prueba de metadatos es necesario validar la marca de tiempo, si cada registro ha asignado la identificación de tiempo.
42	No almacenar metadatos reales	Es muy importante mantener los metadatos reales, porque cualquier pequeño problema con la actualización puede resultar en una cadena de problemas serios.
43	Datos y metadatos no coinciden	Metadatos diferentes a los datos de la bodega.
44	Metadatos no confiables	Los metadatos no confiables e incompletos de las fuentes de datos causan problemas de calidad de datos.
45	No hay reglas para limpieza de metadatos	La falta de reflejo de las reglas establecidas para la limpieza de datos en los metadatos causa una calidad deficiente de los datos.
46	No hay informe de errores	La falta de informes de errores, validación y actualizaciones de metadatos causan problemas de calidad de datos.

Tabla 6. Descripción de riesgos para el dominio de metadatos
Fuente: Elaboración propia.

Evaluación de riesgos

En este capítulo se pretende realizar una evaluación de cada riesgo identificado anteriormente.

La evaluación consiste en calificar en una escala de 1 a 5 cada riesgo por probabilidad e impacto de ocurrencia y así poder identificar los riesgos inherentes.

En la tabla 7 se define la probabilidad de ocurrencia de un riesgo y la tabla 8 define el impacto de un riesgo.

	Probabilidad	**Descripción**
1	Remoto	Muy poco factible que el riesgo se presente
2	Improbable	Poco factible que el riesgo se presente
3	Moderado	Es factible que el riesgo se presente
4	Probable	Posible que el riesgo se presente
5	Casi cierto	Muy posible que el riesgo se presente

Tabla 7. Definición y valor de probabilidad de ocurrencia de un riesgo
Fuente: Elaboración propia.

	Impacto	**Descripción**
1	Bajo	El impacto del riesgo es aceptable y controlable
2	Moderado	El impacto del riesgo es aceptable
3	Intermedio	El impacto del riesgo puede causar daños leves
4	Alto	El impacto del riesgo es peligroso
5	Crítico	El impacto del riesgo es catastrófico

Tabla 8. Definición y valor de impacto del riesgo
Fuente: Elaboración propia

Para dar valores de probabilidad e impacto a cada riesgo se usó el método Delphi.
Según [13] el método Delphi es una herramienta que ayuda a la investigación para obtener datos confiables de un grupo de expertos y en varias ocasiones puede proporcionar solución a problemas complejos.

Para la implementación de este método se deben seguir los siguientes pasos:

1. Identifique un panel de expertos que aporten conocimiento y autoridad, idealmente de una variedad diversa de campos estrechamente relacionados con el tema de investigación, que se comprometan a participar en el estudio.

2. Encuestar a los panelistas individualmente para conocer sus predicciones, prioridades, calificaciones en escalas o cualquier otra pregunta diseñada para capturar de manera más efectiva las respuestas.
3. Agregue los datos recopilados del paso anterior en un informe conciso e inequívoco que identifique las medidas relevantes de tendencia central (por ejemplo, media o mediana) y variación (por ejemplo, desviación estándar) para la distribución del panel.
4. Envíe el informe del paso 3 a cada panelista, con o sin un recordatorio de sus propias respuestas anteriores, y una invitación para volver a responder las mismas preguntas, ahora con la información adicional del informe agregado más reciente.
5. Agregue la nueva ronda de respuestas y compare este informe con la ronda anterior. Si no difieren, el proceso ha convergido y se obtienen los resultados finales, pero si el informe agregado ha evolucionado en la ronda anterior, regrese al paso 3.

A continuación, la tabla 9 muestra cada riesgo identificado en la bodega de datos de CDIAC calificado según los criterios de probabilidad e impacto definidos en la tabla anterior.

	Riesgo	**Probabilidad**	**Impacto**	**Riesgo Inherente**
1	Identificación incorrecta de hechos, dimensiones o relaciones de la bodega	2	3	**6**
2	Relaciones de datos inapropiadas entre tablas	2	4	**8**
3	Pérdida de datos en el proceso ETL	4	4	**16**
4	Valores nulos no encontrados en proceso de ETL	3	4	**12**
5	La imposibilidad de reiniciar el proceso ETL desde los puntos de control sin perder datos	2	4	**8**
6	ETL no eficiente debido a un mal diseño	3	4	**12**

7	Los datos no se cargan completos a la bodega	3	4	**12**
8	Problemas de calidad de los datos en las fuentes de datos	3	4	**12**
9	Selección inadecuada de las fuentes de datos candidatas	2	3	**6**
10	Cambios inesperados en los sistemas de origen	3	4	**12**
11	Uso de diferentes formatos de representación en fuentes de datos	2	2	**4**
12	Presencia de registros duplicados	3	4	**12**
13	Valores faltantes en las fuentes de datos	3	4	**12**
14	Datos mal escritos o en medidas erróneas	2	3	**6**
15	Columnas adicionales	3	2	**6**
16	Valores predeterminados	2	4	**8**
17	No especificar el carácter NULL en fuentes de datos	3	4	**12**
18	Valores atípicos	3	4	**12**

19	Extracción incorrecta de datos	3	4	**12**
20	No actualizar las fuentes de datos	2	4	**8**
22	Actualizaciones no documentadas	2	2	**4**
23	Falta de actualización periódica	3	4	**12**
24	Malas conversiones del sistema	3	5	**15**
25	Columnas que tienen valores de datos incorrectos	2	3	**6**
26	Riesgo semántico	2	3	**6**
27	Corrupción de datos	3	5	**15**
28	Mal uso de herramientas de análisis y minería de datos	3	4	**12**
29	Dificultad en la recopilación de los datos	2	2	**4**
30	No definir los perfiles para los usuarios	3	4	**12**
31	Usuarios con permisos inadecuados	3	4	**12**
32	Usuarios que abusen de los permisos asignados para el desarrollo de su trabajo	3	5	**15**

33	Personal poco capacitado	3	5	**15**
34	Actualización de perfiles	2	3	**6**
35	Inyección SQL	3	5	**15**
36	Truncar tablas	2	5	**10**
37	Espacio lógico	2	5	**10**
38	Espacio Físico	1	3	**3**
39	No tener metadatos	2	4	**8**
40	Valores predeterminados que no se cargan correctamente	2	3	**6**
41	Marca de tiempo no establecida para registros	1	3	**3**
42	No almacenar metadatos reales	2	4	**8**
43	Datos y metadatos no coinciden	2	4	**8**
44	Metadatos no confiables	2	3	**6**
45	No hay reglas para limpieza de metadatos	2	3	**6**

46	No hay informe de errores	4	5	**20**

Tabla 9 evaluación de riesgos
Elaboración propia.

En la tabla 10 se define la matriz de riesgos, los colores nos indican:

Verde: Bajo riesgo

Amarillo: Medio riesgo

Rojo: Alto riesgo

En esta tabla se ubican los riesgos inherentes y los que queden en el sector rojo serán los riesgos a los que se les diseñe controles para este caso en específico.

	IMPACTO					
		BAJO		**MODERADO**	**ALTO**	
	Indicadores	**1**	**2**	**3**	**4**	**5**
PROBABILIDAD	**5**	**5**	**10**	**15**	**20**	**25**
	4	**4**	**8**	**12**	**16**	**20**
	3	**3**	**6**	**9**	**12**	**15**
	2	**2**	**4**	**6**	**8**	**10**
	1	**1**	**2**	**3**	**4**	**5**

Tabla 10. Matriz de riesgos
Elaboración propia.

Diseño de controles

Cada riesgo crítico debe tener al menos un control que permita evitar su ocurrencia o minimizar su impacto, pero esto no quiere decir que la bodega necesite todos los controles que sean definidos, se debe realizar un estudio del estado actual de la bodega para conocer cuáles son los controles que mejor se adecuan a las necesidades del cliente.

	Riesgo	**Riesgo Inherente**	**Control**
3	Pérdida de datos en el proceso ETL	16	Aplicar algoritmos para establecer un sistema de alertas basado en reglas, que monitorice y alerte acerca de inconsistencias durante todo el proceso [9].
24	Malas conversiones del sistema	15	
27	Corrupción de datos	15	Aplicar metodologías de validación de datos teniendo en cuenta tiempo de ejecución, cobertura de datos y eficiencia de las consultas.
32	Usuarios que abusen de los permisos asignados para el desarrollo de su trabajo	15	Realizar seguimiento a las consultas realizadas por los usuarios del sistema.
			Registrar intentos de conexión a la bodega de datos desde aplicaciones ajenas a las autorizadas.
33		15	Optimizar filtros de selección de personal.

	Personal poco capacitado		Realizar cursos o capacitaciones necesarias para el personal encargado del sistema.
35	Inyección SQL	15	Realizar la actualización de la aplicación a la última versión estable disponible que exista
46	No hay informe de errores	20	Realización y validación de informes para cada error o actualización al sistema

Tabla 11. Diseño de controles para riesgos de alto impacto
Elaboración propia.

CAPÍTULO 4

Informe final para la bodega de datos del CDIAC

La realización de esta auditoría se hizo bajo la metodología propuesta en este trabajo de grado, la cual propone, en primer lugar, definir los dominios propios de la bodega, como se hace de manera similar en COBIT 4.0.

La identificación de los riesgos se realiza a partir de cada dominio y para realizar dicha definición se usaron técnicas de auditoría con ayuda de computadora (TAAC). Para la evaluación de riesgos se contó con el conocimiento y apoyo de:

Néstor Darío Duque Méndez, profesor creador de la bodega de datos.

Daniel Andrés Espinoza, actual encargado del funcionamiento y mantenimiento de la bodega de datos.

Emilcy Juliana Hernández Leal: Actual coordinadora de proyectos que trabajan bajo el funcionamiento de la bodega y páginas del CDIAC, encargados al grupo de investigación GAIA.

Luis Felipe Londoño: Estudiante auxiliar encargado de desarrollar el módulo de auditoría de la bodega de datos.

Después de la evaluación de riesgos se ubican los riesgos en la matriz de riesgos y para este trabajo de grado se decide realizar un diseño de controles solo para los riesgos ubicados en la zona roja de la matriz. Después de realizar el anterior proceso, en la tabla 12, se muestran los riesgos críticos identificados para la bodega de datos del Centro de Datos e Indicadores Ambientales de Caldas.

Riesgo	Dominio	R.I	Control
Pérdida de datos en el proceso de ETL	ETL	16	Aplicar algoritmos para establecer un sistema de alertas basado en reglas, que monitorice y alerte acerca de inconsistencias durante todo el proceso [8].
Malas conversiones en el sistema	ETL	15	Utilizar una herramienta de extracción y administración de datos similar a Oracle Warehouse Builder [14].
Usuarios que abusen de los permisos asignados para el desarrollo de su trabajo	Usuarios	15	Realizar seguimiento a las consultas realizadas por los usuarios del sistema.

			Registrar intentos de conexión a la bodega de datos desde aplicaciones ajenas a las autorizadas.
Personal poco capacitado	Usuarios	15	Optimizar filtros de selección de personal. Realizar cursos o capacitaciones necesarias para el personal encargado del sistema.
Corrupción de datos	ETL	15	Aplicar metodologías de validación de datos teniendo en cuenta tiempo de ejecución, cobertura de datos y eficiencia de las consultas.
Inyección SQL	Bodega de datos	15	Realizar la actualización de la aplicación a la última versión estable disponible que exista
No tener informe de errores	Metadatos	20	Realización y validación de informes para cada error o actualización al sistema

Tabla 12. Reporte de errores
Fuente: Elaboración propia.

Como vemos en la tabla anterior en cuatro de los seis dominios definidos se encontraron riesgos críticos, siendo el dominio ETL el más propenso a riesgos seguido por el dominio usuarios, con un total de 25 y 5 riesgos identificados respectivamente, clasificados como se muestra en la tabla 13.

Clasificación del riesgo para el dominio	**ETL**	**Usuarios**
Número de riesgos	**3**	**2**
	22	**2**
	0	**1**

Tabla 13. Clasificación en la matriz de riesgos identificados
Fuente: Elaboración propia

Como conclusión, se tiene que el dominio ETL es el proceso en el cual se debe tener más cuidado, ya que es el dominio con más riesgos identificados, claramente pueden llegar a existir más riesgos para cualquiera de los dominios definidos, por lo cual se recomienda aplicar metodologías y usar herramientas que apoyen la extracción, transformación y carga de datos en la bodega.

Bibliografía

[1] “IDEA,” *CDIAC*. [Online]. Available: http://cdiac.manizales.unal.edu.co/inicio/cdiac.php. [Accessed: 24-Nov-2019].

[2] N. D. Duque Méndez and A. Tamayo Alzate, “Data Warehouse (Bodega de Datos). Herramienta para la toma de decisiones (Parte I) Sistema de Procesamiento de Transacciones (OLTP: On-Iine Transaction Processing),” *Univ. Nac. sede Manizales*, no. 1, p. 126, 2001.

[3] L. Castelán and J. Ocharán, “Diseño de un Almacén de datos basado en Data Warehouse Engineering Process (DWEP) y HEFESTO,” *Univ. Veracruzana*, p. 10, 2012.

[4] ISACA, (2012) *Un Marco de Negocio para el Gobierno y la Gestión de las TI de la Empresa*. Rolling Meadows, pp. 75-77.

[5] GUÍA ISO/IEC 73:2002. (2010). ICS: 01.120 / Normalización. Reglas generales. CTN 1 - Normas Generales.

[6] C. M. Herrera Cruz, S. G. Rivera García, and C. A. Vázquez López, “Auditoría basada en Riesgos a empresas que fabrican y comercializan productos pirotécnicos,” Universidad de El Salvador, 2014.

[7] Rodero, J., Toval, J., Piattini, M. (1999) "The audit of the Data Warehouse Framework". España.

[8] N. D. Duque-Méndez, V. Tabares-Morales, and H. González, “ESIA Expert System for Systems Audit Risk-Based,” in *Advances in Artificial Intelligence - IBERAMIA 2018*, 2018, pp. 483–494.

[9] Singh, R., Singh, K., (2010) “A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing,” *International Journal of Computer Science Issues*, vol. 7, no. 3, pp. 41–50.

[10] "PostgreSQL: About", *Postgresql.org*, 2020. [Online]. Available: https://www.postgresql.org/about/. [Accessed: 13- Feb- 2020].

[11] Tanuska, P., "The Proposal of the Essential Strategies of Data Warehouse Testing," *Central European Conference on Information and Intelligent Systems*, pp. 1-5, 2008. Available: http://ezproxy.unal.edu.co/docview/1314945435?accountid=150292.

[12] García, M, J., “Database Main Threats Analysis Using MySQL Server”.

[13] Flostrand, A., Pitt, L., Bridson, S. (2017). The Delphi technique in forecasting -A 42 year bibliographic analysis (1975-2017). *Technological Forecasting and Social Change*. vol 150. doi: 10.1016/j.techfore.2019.119773.

[14] G. Lumpkin, “Oracle Database 11g para Data Warehousing e Inteligencia de Negocios,” *Inf. Ejec.*, p. 11, 2013.