MODELO PARA EL PROCESO DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA EN BODEGAS DE DATOS. UNA APLICACIÓN CON DATOS AMBIENTALES

MODEL FOR THE EXTRACTION, TRANSFORMATION AND LOAD PROCESS IN DATA WAREHOUSES. AN APPLICATION WITH ENVIRONMENTAL DATA

Néstor Darío Duque Méndez[1], Emilcy Juliana Hernández Leal[2], Ángela María Pérez Zapata[3], Adrián Felipe Arroyave Tabares[4], Daniel Andrés Espinosa[5]

Fecha de recepción: 18 de marzo de 2016
Fecha de revisión: 4 de mayo de 2016
Fecha de aprobación: 31 de mayo de 2016

Referencia: N. D. Duque Méndez, E. J. Hernández Leal, Á. M. Pérez Zapata, A. F. Arroyave Tabares, D. A. Espinosa (2016). Modelo para el proceso de extracción, transformación y carga en bodegas de datos. Una aplicación con datos ambientales. Ciencia e Ingeniería Neogranadina, 26 (2), pp. 95-109, DOI: http://dx.doi.org/10.18359/rcin.1799

RESUMEN

La administración de bodegas de datos o *datawarehouse* requiere de un procesamiento para garantizar la veracidad, integridad y centralización de los datos cuando existen diversas fuentes de información, haciendo necesario utilizar aplicativos especializados para la Extracción, Transformación y Carga de datos (ETL). Estos aplicativos presentan conflictos en su parametrización, carecen de la implementación de filtros de corrección adaptables a las características de los datos

1. Ing. Mecánico, Ph.D en Ingeniería, profesor asociado, director Grupo de Investigación GAIA, Facultad de Administración, Universidad Nacional de Colombia, sede Manizales, Manizales, Colombia, ndduqueme@unal.edu.co
2. Administradora de Sistemas Informáticos, estudiante de Maestría en Ingeniería Administrativa, Facultad de Minas, Universidad Nacional de Colombia, sede Medellín, Medellín, Colombia, ejhernandezle@unal.edu.co
3. Estudiante de Administración de Sistemas Informáticos, Facultad de Administración, Universidad Nacional de Colombia, sede Manizales, Manizales, Colombia, amperezz@unal.edu.co
4. Estudiante de Administración de Sistemas Informáticos, Facultad de Administración, Universidad Nacional de Colombia, sede Manizales, Manizales, Colombia, afarroyavet@unal.edu.co
5. Estudiante de Administración de Sistemas Informáticos, Facultad de Administración, Universidad Nacional de Colombia, sede Manizales, Manizales, Colombia, daespinosag@unal.edu.co

y pueden demandar altos costos para su implementación. En el presente artículo se plantea un modelo genérico que aplica las etapas de ETL y permite realizar seguimiento del proceso al mantener un registro histórico de errores filtrados y calcular indicadores para identificar la calidad en el procesamiento. La validación del modelo fue realizada sobre un caso de estudio con datos ambientales. El modelo demostró obtener resultados satisfactorios. Se plantea realizar más validaciones del modelo, en otros ámbitos, incluyendo nuevos tipos y estructuras de datos.

Palabras clave: bodegas de datos, consistencia, integridad, procesamiento web, procesos ETL.

ABSTRACT

Data warehouse management requires a procedure to ensure the accuracy, completeness, and centralization of data when there are several sources of information, thus making the use of specialized applications for Extraction, Transformation, and Loading of Data -ETL- necessary. These applications have conflicts with the parameterization, lack the implementation of correction filters adaptable to the data characteristics, and can demand high costs for their implementation. In this article, it is presented a generic model that applies the stages of ETL and allows monitoring the process to keep a historical record of errors filtered and to calculate indicators to identify quality in processing. Model validation was performed on a case study with environmental data. The model showed satisfactory results. Finally, it is planned to conduct validations of the model in other areas, including new types and data structures.

Keywords: datawarehouses, consistency, integrity, web processing, ETL process.

INTRODUCCIÓN

Los procesos de extracción, transformación y carga de datos, mejor conocidos como ETL por sus siglas en inglés (*Extract, Transform, Load*), se enmarcan dentro de las actividades clave en el contexto de las bases de datos, ya que por medio de su combinación permiten hacer el traslado de datos de una fuente a otra. Principalmente este término se ha asociado a procesos propios de la construcción de bodegas de datos, o *datawarehouse* [1]. Las bodegas de datos son repositorios de información recolectada de múltiples fuentes, unificada bajo un esquema y que usualmente se encuentra en un mismo lugar [2].

Dentro de las fases de la construcción de un *datawarehouse*, el ETL es una de las tareas con mayor costo, tanto por tiempo como por recursos, estando esta labor asociada a la unificación de datos provenientes de diferentes fuentes, con estructuras y formatos variantes. A pesar de que existen diversas herramientas, tanto libres como propietarias [3-5], para el modelado y ejecución de estos procesos, no siempre es posible alcanzar el nivel de personalización que requieren algunos problemas complejos, donde la variedad de fuentes y esquemas de datos dificultan la labor.

En este artículo se presenta un modelo cuyo objetivo es tener un acercamiento a la

optimización de procesos de ETL y hacerlos más eficientes para la construcción y poblado de bodegas de datos. El modelo incluye un módulo traductor, una fase de filtrado detectivo y correctivo y una migración final de datos; además, se proponen tareas adicionales como administración de la bodega en el tiempo y generación de indicadores de confianza y soporte. Estos indicadores permiten hacer un seguimiento de la calidad de los datos cargados en la bodega. Se hicieron pruebas del modelo en un caso de estudio con datos ambientales que mostraron resultados satisfactorios.

El resto del artículo se estructura de la siguiente manera: en las Secciones 2 y 3 se presentan algunos conceptos básicos y trabajos relacionados, respectivamente. La Sección 4 contiene una descripción de la arquitectura del modelo propuesto, incluyendo las diferentes fases. Por su parte, en la Sección 5 se introduce el caso de estudio, las pruebas realizadas y los resultados obtenidos. Por último, en la Sección 6 se traen a consideración las conclusiones y trabajos futuros.

1. CONCEPTOS BÁSICOS

Uno de los aspectos importantes a considerar en un sistema para la administración de datos es el almacenamiento; dado que de esto dependen muchos factores, como acceso, disponibilidad, escalabilidad, facilidad de recuperación, estructuración de consultas, tiempos de respuesta, entre otras. Existen diferentes técnicas de almacenamiento de datos, van desde los modelos más tradicionales (relacionales) hasta las nuevas tendencias que incluyen la posibilidad de almacenar datos no estructurados o semiestructurados [6]. Sin embargo, en los dos casos se requiere de un tratamiento previo de los datos, que ayude a garantizar la consistencia de los mismos. Dicho tratamiento enmarca los procesos de ETL, los cuales se asocian principalmente a los proyectos de *datawarehouse*. Un *datawarehouse,* o bodega de datos, es un repositorio de información recolectada de múltiples fuentes, unificada bajo un esquema y que usualmente se encuentra en un mismo lugar [7]. Los *datawarehouse* están construidos por medio de un proceso de limpieza, integración, transformación, carga y actualización periódica de datos (ver Figura 1).

Asimismo, los *datawarehouse* están modelados usualmente por una estructura de base de datos multidimensional, donde se tiene una serie de atributos agrupados en unas dimensiones que, a su vez, hacen parte de un esquema [8]. Las bodegas de datos son utilizadas en diversos campos y organizaciones, ya que el crecimiento del volumen de datos es generalizado y se ha convertido en una tendencia desde los noventa [9]. Existen varios aspectos fundamentales a la hora de caracterizar un *datawarehouse*, como son la orientación a un tema específico, la integración, la no volatilidad y la variación en el tiempo [10].

El hecho de que un *datawarehouse* sea integrado, implica que va a ser alimentado con datos provenientes de diferentes fuentes, los cuales deberán ser limpiados y estructurados bajo un esquema. Ahora bien, cuando se habla de que un *datawarehouse* debe ser no volátil, implica que, contrario a como pasa en los sistemas transaccionales tradicionales (donde se inserta y modifica información de forma constante), en un *datawarehouse* los datos se cargan y acceden generalmente de forma masiva sin ser modificados.

Las arquitecturas utilizadas en los *datawarehouse* son relacional y

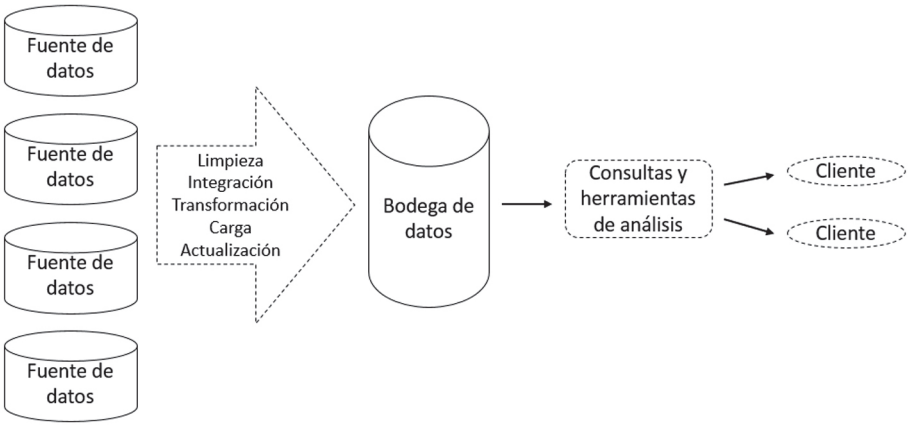

Figura 1. Marco de trabajo típico para la construcción de un datawarehouse.

Fuente: Adaptado de Han, Kamber y Pei, 2011 [5]

multidimensional, la segunda presenta dos estructuras principales, estructura en estrella y estructura copo de nieve. La estructura en estrella es el modelo multidimensional clásico, con una única tabla de hechos rodeada de dos o más tablas de dimensiones. Por su parte, el copo de nieve es una variante que presenta varias tablas de hechos que comparten algunas tablas de dimensiones entre sí [11].

Como se mencionó anteriormente, los procesos de ETL hacen parte fundamental en acciones que van desde la migración sencilla de una base de datos hasta unas mucho más complejas, como la construcción de un *datawarehouse*. Estos se desarrollan en varias fases o actividades, las cuales son definidas en [12] de la siguiente manera. Tarea 1: identificación de las fuentes de datos de las cuales se hará la extracción, suelen ser heterogéneas. Tarea 2: transformación de las fuentes, después de extraer los datos estos pueden ser transformados y generar datos derivados; en esta tarea se suelen hacer tareas como filtrado, conversión, cálculo de valores derivados, generación de llaves, entre otros. Tarea 3: unión de las fuentes, consiste en llevar a un solo almacén diversas fuentes. Tarea 4: seleccionar el destino para cargar los datos. Tarea 5: unión de los atributos de las fuentes con los atributos previamente almacenados en el destino. Tarea 6: carga de datos, comprende el poblado de la bodega de datos con los datos ya limpios y transformados.

2. TRABAJOS RELACIONADOS

A continuación se presentan algunos trabajos relacionados con la formulación conceptual de modelos de ETL y algunos casos de implementación, los cuales destacan la importancia de estos procesos en la construcción de los almacenes de datos. Se destaca también la existencia de algunas brechas por cubrir, principalmente cuando se cuentan con datos de diferente naturaleza (heterogeneidad), y cuando estos presentan vacíos que se deben suplir.

Por su parte, en [13] se concretan en la definición conceptual del proceso de ETL; en específico, hacen hincapié en la importancia de la definición de los procesos de ETL dentro del proceso general de construcción de los *datawarehouse*. Los autores proponen un esquema para el modelado gráfico de las actividades de ETL, dando parámetros para la representación conceptual de estas. En especial, se hace una propuesta para la customización de las relaciones entre los atributos y las respectivas actividades de ETL en cada una de las fases de un proyecto de *datawarehouse*. El trabajo no presenta una aplicación concreta del modelo sobre datos reales, pero es un buen aporte a la hora de hacer la esquematización conceptual de estos procesos.

En [14] se menciona el desarrollo de un modelo conceptual para un esquema de almacenamiento de series de datos hidroclimatológicas a través de un *datawarehouse*. Es de resaltar que en este trabajo se ratifica la importancia de contar con datos limpios y validados, que hayan sido sometidos a una fase de ETL previa a la realización de análisis y generación de conocimiento. Esto con el fin de garantizar la coherencia en cuanto a unidades y periodicidad, y conseguir datos con integridad y consistencia.

En [15] se propone un esquema conceptual para la realización del proceso de ETL, los autores resaltan la importancia de contar con un modelo conceptual estándar para simplificar la representación del proceso de ETL. Se define a *novel conceptual model entity mapping diagram* (EMD por sus siglas en inglés). Este modelo incluye dos capas. Una primera de abstracción, en la cual se presentan cinco objetos: funciones, contenedor de datos, entidades, relaciones y atributos. Los objetos dentro de esta capa de abstracción son una vista de alto nivel, que pueden ser usados para diagramar un escenario EMD. La segunda capa es la *template*, la cual es una expansión de la capa de abstracción. Cada usuario puede diseñar su propio escenario de ETL, incluso añadiendo capas. Construyeron también un *framework* para usar el modelo, el cual consiste en un componente para los recursos de datos, otro componente para el esquema de *datawarehouse* y un último componente de mapeo. El modelo no ha sido testeado en un caso real, esto es planteado como trabajo futuro.

En [16] proponen un enfoque de ETL donde se aplican tablas virtuales para realizar la etapa de transformación antes de la de extracción y de carga. El enfoque es llamado TEL, ya que se cambian las siglas en inglés a *Transform-Extract-Load*. Según los autores, el enfoque reduce la carga de transmisión de los datos y mejora el rendimiento de las consultas por medio de capas de acceso; además muestra el enfoque como factible y práctico.

En [17] se presenta la implementación de un sistema multiagente (SMA) para la realización del proceso de ETL, en el cual se considera la heterogeneidad y disponibilidad de los datos a la hora de crear un almacén de datos. La propuesta presentada por los autores parte de la recopilación de las fortalezas de otros enfoques como los *wrappers* y soluciones *ad-hoc*. El modelo es validado por medio de datos reales y simulados. Se aplicaron algunos indicadores para medir la efectividad y viabilidad técnica de la propuesta, y se obtuvieron resultados satisfactorios; sin embargo, el volumen de datos empleado para las pruebas fue pequeño y no se describe con claridad si hay filtrado de datos.

De la revisión de literatura se concluye que existen diversos modelos y enfoques

conceptuales que permiten esquematizar procesos de ETL, se identifica que existen espacios para trabajar en el marco de la aplicación de modelos concretos y en casos de estudio que tomen datos reales, donde se pueda detectar de forma práctica la problemática asociada a esta fase fundamental dentro de la construcción de los *datawarehouse*.

3. MODELO PROPUESTO

En este trabajo se expone el desarrollo de un modelo para la realización del proceso de extracción, transformación y carga aplicado a una bodega de datos. El modelo, además de incluir la posibilidad de tomar diferentes fuentes de datos, también cuenta con la capacidad de hacer un filtrado, tanto de detección como de corrección, y garantizar con esto la integridad y consistencia de los datos que se almacenan en la bodega de datos.

El modelo concebido es presentado en la Figura 2. Este modelo está desarrollado para desempeñarse en un entorno web. A continuación se describe en detalle cada uno de los componentes que lo estructuran, y se revisa la figura de izquierda-derecha.

3.1 Fase de prerrequisitos (extracción)

Los datos pueden estar representados en diferentes formas de almacenamiento, como por ejemplo archivos planos y repositorios de datos estructurados (bases de datos), los cuales pueden presentar problemas y errores en su adquisición. Por lo anterior es necesario contar con un proceso previo que se ha llamado 'traducción'; este proceso consiste en la toma y estandarización de la estructura de los datos, para luego proceder a la entrega de los mismos a la siguiente fase del proceso.

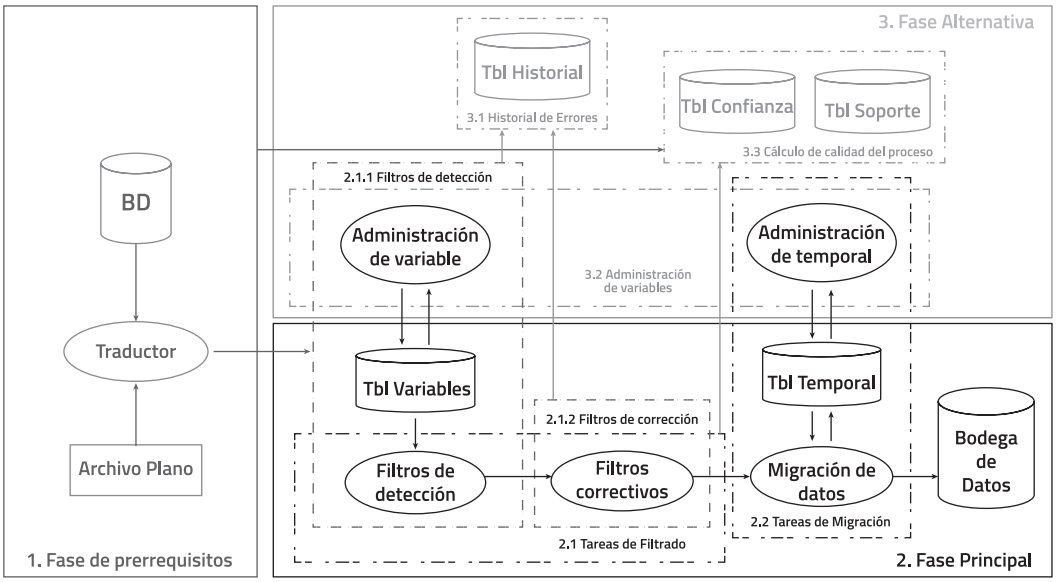

Figura 2. Modelo de ETL propuesto.

Fuente: Elaboración propia.

3.2 Fase principal

Consta de filtrado y migración, es la encargada de recibir los datos organizados en una estructura estándar, construida en la fase de traducción, y a partir de esto realiza la tarea de filtrado; luego, si es el caso, hacer las correcciones predefinidas y, finalmente, migrar a la bodega de datos. Este proceso está soportado en una tabla temporal (tbl-temporal), la cual presenta la misma estructura de la tabla a la cual se va a migrar. Después de ingresar los datos en la tabla temporal, la fase principal de filtrado y migración se divide en dos tareas consecutivas, las cuales se describen a continuación.

3.2.1 Tarea de filtrado

Está compuesta por dos actividades, las cuales son: filtrado de detección y filtrado de corrección de fallas. En la primera se reciben los datos originales en una estructura estándar, se examinan y detectan posibles errores. En la segunda se aplica el proceso de corrección correspondiente.

- *Filtrado de detección:* en esta actividad se detectan los errores presentes en las mediciones de cada una de las variables; para ello se emplea una tabla de variables (tbl-variables), la cual contiene los filtros y restricciones para cada variable entrante en el modelo. Estos filtros y restricciones son determinados por profesionales con dominio en el ámbito de aplicación del modelo. Es la encargada de detectar los valores atípicos e inconsistencias en los datos fuente.
- *Filtrado de corrección de fallas:* en esta actividad se reciben los datos con los errores detectados y organizados y se sigue un estándar específico. Los filtros de corrección deben estar determinados por profesionales en el área de aplicación del modelo. Después de aplicar las acciones correctivas correspondientes a las mediciones de cada variable, los datos están listos para la siguiente tarea.

3.2.2 Tarea de migración de datos

Esta es la encargada de cargar los datos de la tabla central (tbl-temporal) a la bodega de datos.

3.3 Fase alternativa

Adicional a las fases propias del proceso de ETL, se definen actividades importantes referentes a la evaluación del proceso, así como de su administración en el tiempo.

3.3.1 Historial de errores

Esta actividad es la encargada de almacenar la información referente a los errores, tal como: descripción del error, valor corregido, posición donde se encuentra el error, tipo de corrección aplicado, estación, variable, fecha y hora en que fue calculado el error.

3.3.2 Administración de variables

Esta actividad es la encargada de proporcionar los CRUD (*Create, Read, Update y Delete*, por sus siglas en inglés), para las tablas temporales, las cuales contienen los datos necesarios para administrar los filtros; dicha actividad es desarrollada con el objetivo de hacer el proceso de administración en el tiempo y facilitar la manutención y extensión del proceso.

3.3.3 Cálculo de calidad del proceso

A continuación se describen los indicadores diseñados para determinar la calidad de los datos cargados a la bodega de datos. Estos

indicadores permiten a los administradores de la bodega tener la percepción de los datos fuente que son corregidos y los datos que se conservan sin modificación alguna.

- *Confianza (1):* representa la relación entre la cantidad de registros que ingresan sin errores y el total de registros entrantes por cada variable filtrada; este indicador se calcula con el objetivo de obtener información relevante en cuanto a la calidad real de los datos por cada variable.

$$\text{Confianza} = \frac{\textit{Total registros válidos}}{\textit{Total registros entrantes}} \quad (1)$$

- *Soporte (2):* representa la relación entre la cantidad de registros que ingresan sin errores y el total teórico de medición por unidad de tiempo (día, semana, mes...), este indicador representa la calidad teórica de los datos.

$$\text{Soporte} = \frac{\textit{Total registros válidos}}{\textit{Valor teórico de la medición}} \quad (2)$$

Nota: Cabe destacar que de no calcularse la confianza, el soporte y el historial de errores al momento de ejecutar el proceso de ETL, estos importantes datos no podrán calcularse después.

4. CASO DE ESTUDIO, PRUEBAS Y RESULTADOS

En esta sección se presenta el caso de estudio utilizado para probar el modelo, algunas pruebas realizadas y los resultados obtenidos en estas.

4.1 Caso de estudio

Para la validación del modelo se implementó un caso de estudio real enmarcado en el dominio de los datos ambientales. Actualmente en el departamento de Caldas se cuenta con más de 80 estaciones de monitoreo ambiental, distribuidas en diferentes redes, las cuales están en crecimiento continuo. Dichas estaciones realizan mediciones de variables meteorológicas e hidrométricas, según se disponga en su diseño e instalación. Las estaciones meteorológicas comprenden la medición de las variables precipitación, temperatura del aire, brillo y radiación solar, humedad relativa, velocidad y dirección del viento, presión barométrica y evapotranspiración. Las variables hidrométricas corresponden a registros de nivel y caudal, tomadas en ríos y otras fuentes hídricas.

La captura de los datos es obtenida a través de sensores especiales que son controlados por un PLC –Controlador Lógico Programable–. Este recibe la señal, que es transmitida por los sensores para almacenar su valor numérico equivalente, pero en ocasiones las lecturas pueden generar valores erróneos. Usualmente los valores erróneos se identifican por tomar valores alfabéticos o encontrarse fuera de un rango válido, el cual se establece según el tipo de variable y la posición geográfica de la estación.

La transmisión de los datos se hace siguiendo dos metodologías. El primer método es realizado manualmente, donde se requiere la intervención del operador de la red. Estas estaciones aportan por cada variable un dato diario. El operador debe desplazarse hasta la estación para recolectar los datos medidos y registrarlos en archivos planos con diferentes

estructuras. La segunda forma es automática, por medio de telemetría. Las estaciones realizan la captura y transmisión de los datos en un intervalo de tiempo aproximado de cinco minutos. Estos son almacenados en bases de datos localizadas en diversos servidores de adquisición. Las bases de datos contienen varias tablas, donde se establece una correspondencia con cada estación, además cuentan con estructuras de diversa naturaleza.

En ambos casos, la transmisión de los datos proporciona diversas estructuras que son definidas según los requisitos de las variables medidas y el estándar establecido por la organización propietaria de la estación. En el caso de la operación manual, donde se producen archivos planos, estos pueden variar el orden en el que se citan las variables en sus columnas.

El contexto de los procesos ETL proporciona la facilidad de filtrar y centralizar los datos ambientales del departamento de Caldas. Este proceso permite garantizar el tratamiento adecuado de los datos, puesto que en el modelo se contempla la estandarización, la eliminación o modificación de información inválida generada por ruidos durante la captura de la medición y su carga en una única localización.

En este caso de estudio, se presentan diversas fuentes de datos, las cuales no manejan un estándar general en su estructura. Los filtros de detección aplicados (ver Tabla 1) permiten encontrar datos atípicos, y los filtros de corrección (ver Tabla 2) permiten completar algunos datos faltantes o atípicos. Los parámetros de ambos filtros han sido determinados según el conocimiento de expertos en los temas ambientales, además fue necesario que dichos expertos realizarán un estudio exhaustivo de los datos para identificar los patrones.

Luego de este proceso, es necesario efectuar un volcado de los datos para conservar una única estructura, la cual es presentada en [14]. La estructura de la bodega fue concebida con el objetivo de tener un almacenamiento eficiente de los datos, que garantice un tratamiento eficaz previo a la investigación meteorológica e hidrológica y que, a su vez, permita cubrir la oportunidad que se vislumbra en el análisis de variables hidroclimatológicas; esta corresponde a un modelo multidimensional de estrella centralizada, este diseño fue adoptado con el fin de permitir diferentes niveles de granularidad en las búsquedas que se efectúen para la extracción y futuro procesamiento de los datos. Para el modelo en estrella centralizado se tiene la tabla de hechos que almacena la información de las mediciones tomadas en las estaciones, y tres tablas de dimensiones, que hacen referencia a información propia de la estación, fecha y tiempo de la medición; las dimensiones mantienen una relación con la tabla de hechos a través de sus identificadores.

Al implementar el modelo presentado en este documento, la etapa inicial del proceso contempla la posibilidad de reestructurar los datos a través del traductor, el cual recibe los existentes desde diversas fuentes con estructuras variadas y realiza una revisión del origen de llegada de los datos con el fin de identificar las llaves subrogadas de los mismos y así mantener la información sobre la estación, la fecha y el momento del día en el que se efectuó la medición. Es importante resaltar que la implementación de este modelo ha sido desarrollada de tal manera que sea completamente compatible con la estructura de la base de datos que almacena los datos ambientales.

Tabla 1. Ejemplo de filtros de detección aplicados.

Estación ambiental	Variables climatológicas		
	Presión barométrica (mmHg)	Temperatura del aire (°C)	Humedad relativa (%)
Bosques del Norte	Entre 585 – 602	Entre 5 y 30	Entre 20 – 100
Alcázares	Entre 595 – 612	Entre 5 y 32	
Aranjuez	Entre 600 – 620		
Carmen	Entre 590 – 602		
Emas	Entre 585 – 610		
Hospital de Caldas	Entre 585 – 604		
Ingeominas	Entre 585 – 596		
La Palma	Entre 598 – 620		
Niza	Entre 580 – 600		
Posgrados	Entre 585 – 600		
Yarumos	Entre 580 – 600		
Enea	Entre 595 – 610	Entre 5 y 33	
Chec-Uribe	Entre 605 – 620	Entre 8 y 32	Entre 30-100
Nota: los rangos trabajados fueron suministrados por expertos en monitoreo de variables hidrometeorológicas y con conocimiento del clima de la región. Estos valores son determinados después de años de monitoreo constante y análisis de los datos.			

Fuente: Elaboración propia.

En la segunda fase, al aplicar las tareas de filtrado y cargue de datos, se aplica la identificación y corrección de las mediciones ambientales con el fin de mantener veracidad en los datos almacenados. Cuando se aplica este proceso, de forma paralela se genera un reporte, donde es registrado un histórico sobre los datos filtrados y los datos corregidos, lo que aumenta las probabilidades de que el proyecto ambiental sea auditable, garantizando la integridad de la información extraída del sistema en general. Adicional a la generación del reporte histórico de las condiciones de llegada y salida de los datos en el proceso ETL, se ha requerido en el proyecto aplicar indicadores que permitan definir la eficiencia del sistema de filtrado de datos en las mediciones ambientales, estos indicadores permiten a los usuarios que consultan los datos ambientales directamente en la bodega de datos, identificar una confiabilidad (E1) sobre la cantidad de datos que fueron necesarios aplicar a los filtros

Tabla 2. Ejemplo de filtros de corrección aplicados.

Variable	Operación para el reemplazo	Opción alterna
Temperatura	Promedio	No aplica
Presión barométrica	Promedio	No aplica
Humedad relativa	Promedio	No aplica
Radiación solar	Promedio	No aplica
Precipitación	Mirar el dato anterior y posterior, si hay variación verificar que sea igual a 0,2 y colocar el mismo valor del dato anterior al error	Si la variación es mayor a 0,2 dejar "-", no hacer cambio, pero en el valor posterior al error agregar un campo de observación para poner "Ppt acumulada"
Evapotranspiración	Mirar el dato anterior y posterior al error, si son iguales colocar el mismo valor, si son diferentes colocar el dato anterior	No aplica

Fuente: Elaboración propia.

de corrección, y un indicador de soporte (E2) que permita identificar si la estación transmitió la cantidad de datos que debía transmitir según el periodo de tiempo estipulado, debido a que en algunas ocasiones las estaciones pueden ser suspendidas temporalmente por tareas de mantenimiento preventivo o correctivo. El sistema basado en el modelo finaliza con la etapa de volcado de datos, en la cual se cargan los tratados a la bodega de datos del caso de estudio.

4.2 Pruebas y resultados

Para validar el funcionamiento del modelo de ETL propuesto, se realizó la gestión de un conjunto de datos medidos en cada una de las diferentes estaciones situadas en el departamento de Caldas. Esta gestión incluye la identificación de las diferentes estructuras recibidas, que corresponden a las diferentes fuentes de datos.

Como resultado del procesamiento de los datos ambientales tomados de los servidores donde reposan las mediciones de cada red, se recolectaron un total de 7'465.184 datos del periodo comprendido desde enero de 2012 hasta enero de 2016. El proceso tardó un total de 68 horas y 32 minutos en el filtrado y carga, de manera continua. Con lo anterior se tiene que cada minuto fueron filtrados y cargados aproximadamente 109 datos. Al iniciar las pruebas, la bodega contaba con 12'912.708 datos, logrando llegar a un total de 20'377.892 datos después del proceso.

Se tomó una muestra de datos durante el tiempo de pruebas del modelo (Tabla 3), con el fin de conocer el comportamiento de filtrado en un momento determinado del día.

Tabla 3. Muestra de datos filtrados en serie de tiempo.

i	Hora	X (Datos cargados por hora)	σ^2= (Xi - $\overline{X}$)^2	σ
1	8:00 - 9:00	166255	883100089	29717
2	9:00 - 10:00	115461	444239929	21077
3	10:00 - 11:00	125298	126337600	11240
4	11:00 - 12:00	139135	6744409	2597
			Σ (σ^2) / 4 = 365105506,75	σ = $\sqrt{}$365105506,75 = **19107,73**

Fuente: Elaboración propia.

Con la ejecución de un análisis estadístico [18] sobre una muestra de datos, se obtuvo un promedio de 136.538 datos filtrados y cargados por hora y una desviación estándar de 19.108 datos cada vez. La media indica que el 50% de los datos filtrados han tenido un rendimiento inferior a 136.538 datos por minuto, así como el otro 50% un rendimiento superior al mismo número de datos mencionado, por minuto. Hay que recordar que la precisión de la media se evalúa con la varianza y da una cuantía de la dispersión general. La desviación estándar es igual a 19.107,73 datos, este valor indica que en promedio los valores de la variable se desvían 19.108 datos de la media. Es decir, la distancia promedio a la que se sitúan los valores respecto de la media.

La variabilidad de la cantidad de datos filtrados por hora es de 13,99%, lo que muestra que el promedio es representativo, es decir los datos son homogéneos. La muestra arroja resultados positivos, sin embargo, se pudo detectar que el número de datos filtrados depende de la estación y la hora en que se evalúe; ya que se presentaron espacios de tiempo con sobrecarga en el servidor, disminuyendo con esto el rendimiento del proceso. Se considera que las mejores horas para el procesamiento corresponden a la madrugada (00:00 a 7:00 h).

Al recibir los datos se detectaron inconvenientes con los nombres que presentaban dichos datos, puesto que no eran uniformes, lo cual llevó a la necesidad de aplicar un proceso de conversión para que fueran homogéneos en su interpretación. Como parte de la solución se realizó lectura porcionada de texto, creando una nueva columna dentro de la tabla variables para registrar las formas alternas en que llegan los datos para su identificación.

El tiempo de respuesta se vio afectado por una sobrecarga en el proceso. Al evaluar los datos de las bases de datos, por lo general tardaba más de 10 minutos, lo que provocaba que el servidor se interrumpiera al sobrepasar este tiempo y al no encontrar respuesta reiniciaba el proceso. Se decidió hacer un análisis por lotes, evaluando porciones de 1.000 registros cada vez (se probó previamente sobre la bodega de prueba que el análisis de 1.000 datos tardará menos de 10 minutos), se debía tomar una

cantidad de datos que tardara menos de los 10 minutos en el proceso de filtración y cargue, pero que fuera una cantidad no tan pequeña (no menor a 1.000) para que el servidor no se saturara recargando la página seguidamente, esta cantidad de datos fue determinada por las pruebas que se iban realizando cada vez.

Asi mismo, se realizó un análisis de confiabilidad, donde se detectó que los datos presentan un promedio de 0,8181 con una desviación estándar de 0,1397 (Tabla 4). De lo cual, según la regla empírica [19], se puede inferir que:

- ✓ Los datos siguen una distribución normal con sesgo debido a que la desviación estándar es muy pequeña.
- ✓ Se puede inferir que entre el 0,6784 y el 0,9578 se encuentran el 68% de los datos.
- ✓ Se ratifica el sesgo y se determina su dirección hacia 1. Debido a que para las dos y tres desviaciones sobrepasa el tope máximo del valor. Lo cual representa valores atípicos fuera de las 4 desviaciones hacia el cero, una característica importante de los datos a examinar.
- ✓ Cabe destacar que el porcentaje de valores que pueden exceder las tres desviaciones según la regla empírica es de 0,3%, lo cual representa una anomalía en la distribución para los datos cercanos al cero.

Para el análisis de historial de errores se hizo un conteo de la cantidad de errores por variable y se obtuvo el error más común por la misma. Después se obtuvo el promedio de errores globales por variable, así como el error más común global por variable. Se detectó un promedio de 565.812 errores por variable y el error más común hallado fue "Dato no válido". También se detectó que la variable con más errores es "Radiación solar".

Tabla 4. Análisis de confianza, soporte e historial corrección

Variable	Confianza				Soporte				Historial corrección	
	Prom.	Desv.	Max.	Min.	Prom.	Desv.	Max.	Min.	Cantidad errores	Error más común
Radiación solar	0,93	0,11	1,00	0,00	0,95	0,15	8,75	0,00	1.268.021	Dato no válido
Evapotranspiración	0,94	0,10	1,00	0,00	0,96	0,14	8,75	0,00	0	-
Presión barométrica	0,92	0,12	1,00	0,00	0,95	0,16	8,75	0,00	523.766	Dato no válido
Dirección viento	0,93	0,11	1,00	0,00	0,95	0,15	8,75	0,00	480.274	Dato no válido
Velocidad viento	0,93	0,12	1,00	0,00	0,95	0,16	8,75	0,00	500.885	Dato no válido
Caudal	0,90	0,16	1,00	0,00	0,92	0,19	8,75	0,00	549.033	Dato no válido
Nivel	0,83	0,26	1,00	0,00	0,85	0,29	8,75	0,00	1.123.359	Dato no válido
Humedad relativa	0,91	0,14	1,00	0,00	0,94	0,17	8,75	0,00	597.961	Dato entrante erróneo
Brillo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0	-
Temperatura	0,81	0,26	1,00	0,00	0,82	0,28	8,45	0,00	1.180.638	Dato entrante erróneo
Precipitación	0,90	0,15	1,00	0,00	0,93	0,19	8,75	0,00	0	-
Promedio	0,82	0,14	0,91	0,00	0,84	0,17	7,93	0,00	565.812	Dato no válido

Fuente: Elaboración propia.

5. CONCLUSIONES

Se destaca la complejidad de las tareas de ETL dentro del proceso de construcción de un *datawarehouse*, representada tanto en el costo como en el consumo de tiempo y recursos. En este artículo se presentó el desarrollo conceptual, implementación y pruebas de un modelo de ETL para datos hidrometeorológicos, el cual fue probado con fuentes de datos reales y de diversa naturaleza. El caso de estudio incluyó un volumen considerable de datos y se lograron conseguir buenos resultados.

Cabe resaltar que al utilizar datos reales, el ejercicio realizado contribuye tanto en el ámbito investigativo como en la aplicación del conocimiento en un caso real y de alta relevancia, existiendo un alto interés por parte de diferentes entes y organizaciones en el ámbito ambiental. Aunque se haya utilizado un caso de aplicación concreto, se puede aplicar el modelo en otros dominios con facilidad, basta con seguir los pasos, definir los filtros particulares que se deseen aplicar e identificar las fuentes de datos que alimentarán el modelo.

Como trabajo futuro se plantea aplicar el modelo ETL para hacer tratamiento de otro tipo de datos ambientales, correspondientes a mediciones de calidad del aire, donde se tratan: dióxido de azufre (SO_2), ozono (O_3), dióxido de carbono (CO_2) y materiales particulados de 10 y 2,5 micras (PM10 - PM2,5). También se plantea hacer pruebas en otros dominios, particularmente con datos educativos.

AGRADECIMIENTOS

El trabajo presentado en este artículo fue financiado parcialmente por el programa de la Facultad de Administración de la Universidad Nacional de Colombia, sede Manizales, titulado: "Fortalecimiento de las capacidades de análisis de datos en ambientes educativos que permitan procesos de adaptación", con código 20278.

REFERENCIAS

[1] Kimball, R., Ross, M., Thornthwaite, W., Mundy, J. and Becker, B. (2008). *The Data Warehouse Lifecycle Toolkit*. Indianapolis, USA: Wiley Publishing, Inc.

[2] Calabria-Sarmiento, C. J. (2011). Construcción y poblamiento de un datawarehouse basado en el paradigma de bases de datos objeto relacional. *Prospect*, 9(1), pp. 69-77.

[3] Talend (2016). *Application Integration. The best way to accelerate delivery of real-time application integration*. En: http://www.talend.com/products/application-integration (enero de 2016).

[4] Pentaho (2016). *Data Integration. Pentaho Community.* En: http://community.pentaho.com/projects/data-integration/ (enero de 2016).

[5] CloverETL (2016). *CloverETL Rapid Data Integration*. En: http://www.cloveretl.com/products/community-edition (enero de 2016).

[6] Jaramillo Valbuena, S. y Londoño, J. M. (2015). Sistemas para almacenar grandes volúmenes de datos. *Revista Gerencia Tecnológica Informática*, 13(37), pp. 17-28.

[7] Van den Hoven, J. (1998). Data Warehousing: Bringing it All Together.

Information Systems Management, 15(2), pp. 92-96. doi: 10.1201/1078/43184.15.2.19980301/31127.16

[8] Han, J., Kamber, M. & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Waltham, MA, USA: Elsevier. Tercera edición.

[9] Chaudhuri, S. & Dayal, U. (1997). An Overview of Data Warehousing and OLAP Technology. *ACM SIGMOD Record*, 26(1), pp. 65-74. doi: 10.1145/248603.248616

[10] Shi, D., Lee, Y., Duan, X. & Wu, Q. H. (2001). Power system data warehouses. *IEEE Computer Applications in Power*, 14(3), pp. 49-55. doi: 10.1109/mcap.2001.952937

[11] Tamayo, M. & Moreno, F. J. (2006). Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP. *Ingeniería e Investigación*, 26(3), pp. 135-142.

[12] Trujillo, J. & Luján-Mora, S. (2003). A UML Based Approach for Modeling ETL Processes in Data Warehouses. En I.-Y. Song, S. W. Liddle, T.-W. Ling y P. Scheuermann, *Conceptual Modeling - ER (2003)*, Berlin Heidelberg: Eds. Springer, pp. 307-320. doi: 10.1007/978-3-540-39648-2_25

[13] Vassiliadis, P., Simitsis, A. & Skiadopoulos, S. (2002). Conceptual Modeling for ETL Processes. En: *Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP*, New York, NY, USA, pp. 14-21. doi: 10.1145/583890.583893

[14] Duque-Méndez, N. D., Orozco-Alzate, M. & Vélez, J. J. (2014). Hydro-meteorological data analysis using OLAP techniques. *Revista DYNA*, 81(185), pp. 160-167. doi: 10.15446/dyna.v81n185.37700

[15] El-Sappagh, S. H. A., Hendawi, A. M. A. & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University - Computer and Information Sciences,* 23(2), pp. 91-104. doi: 10.1016/j.jksuci.2011.05.005

[16] Guo, S. S., Yuan, Z. M., Sun, A. B. & Yue, Q. (2015). A New ETL Approach Based on Data Virtualization. *Journal of Computer Science and Technology*, 30(2), pp. 311-323. doi: 10.1007/s11390-015-1524-3

[17] Betancur-Calderón, D. & Moreno-Cadavid, J. (2012). Una aproximación multi-agente para el soporte al proceso de extracción-transformación-carga en bodegas de datos. *Revista Tecno Lógicas*, 28, pp. 89-107.

[18] Morales, A. E. (2012). *Estadística y probabilidad*. Chile.

[19] Johnson, R. & Kuby, P. (2012). *Estadística elemental*. México, D.F.: Cengage Learning. 11° edición, pp. 95-102.