Ingeniería y Competitividad
Vol 26 (3) 2024; doi: 10.25100/iyc.v26i3.13644

Vol 26 (3)

ISSN
0123-3033
e- 2027-8284

Artículo Original

Uso de Arquitectura Dirigida por Modelos en el almacenamiento de datos de PM 2.5 y salud pública

Use of Model-Driven Architecture in the storage of PM 2.5 and public health data

James A. Vergara-Correa[1] Jorge E. Giraldo-Plaza[1] Miriam Gómez-Marin[1] Juan P. Holguín-Marin[2] Nora A. Montealegre-Hernández[3] Juan G. Piñeros-Jiménez[3]

[1]Politécnico Colombiano Jaime Isaza Cadavid. Medellín, Colombia.
[2]Universidad Nacional de Colombia. Medellín, Colombia.
[3]Universidad de Antioquia. Medellín, Colombia.

Resumen

Introducción: en este artículo, se aborda el almacenamiento de datos sobre eventos de salud y partículas PM2.5 en la ciudad de Medellín, Colombia. La consolidación de datos provenientes de fuentes heterogéneas representa un desafío significativo en este contexto.
Objetivo: el objetivo de este estudio es proponer un metamodelo que facilite la integración y almacenamiento de estos datos, utilizando un enfoque basado en modelos.
Métodos: se desarrolló un enfoque modelado que identifica aspectos comunes para la construcción de un data warehouse. Se definió una capa de abstracción sobre los modelos conceptuales de materia particulada y eventos de salud.
Resultados: como resultado principal, se obtuvo un prototipo de data warehouse que permite la consolidación eficiente de datos sobre PM2.5 y eventos de salud. Este prototipo demuestra la efectividad del enfoque propuesto en la integración de datos.
Conclusión: se concluye que el uso de un enfoque basado en modelos fortalece la toma de decisiones en políticas de salud pública y estrategias de gestión de calidad en el ámbito sanitario.

Palabras clave: metamodelo, Bodega de datos, Arquitectura, Calidad de aire, Salud pública.

Abstract

Introduction: this paper addresses the storage of data on health events and PM2.5 particles in the city of Medellín, Colombia. The consolidation of data from heterogeneous sources poses a significant challenge in this context.
Objective: the aim of this study is to propose a metamodel that facilitates the integration and storage of these data using a model-based approach.
Methods: a modeled approach was developed to identify common aspects for building a data warehouse. An abstraction layer was defined over the conceptual models of particulate matter and health events.
Results: the main result was the creation of a data warehouse prototype that allows for the efficient consolidation of data on PM2.5 and health events. This prototype demonstrates the effectiveness of the proposed approach in data integration.
Conclusion: it is concluded that using a model-based approach strengthens decision-making in public health policies and quality management strategies in the healthcare sector.

Keywords: metamodel, Data warehouse, Architecture, Air quality, Public health.

¿Cómo citar?

Vergara-Correa, J.A., Giraldo-Plaza, J.E., Gómez-Marin, M., Holguín-Marin, J.P., Montealegre-Hernández, N.A., Piñeros-Jiménez, J.G. Uso de Arquitectura Dirigida por Modelos en el almacenamiento de datos de PM 2.5 y salud pública. Ingeniería y Competitividad, 2024, 26(3)e-20513644

https://doi.org/10.25100/iyc.v26i3.13644

Recibido: 20-03-24
Evaluado: 16-05-24
Aceptado: 15-08-24
Online: 12-09-24

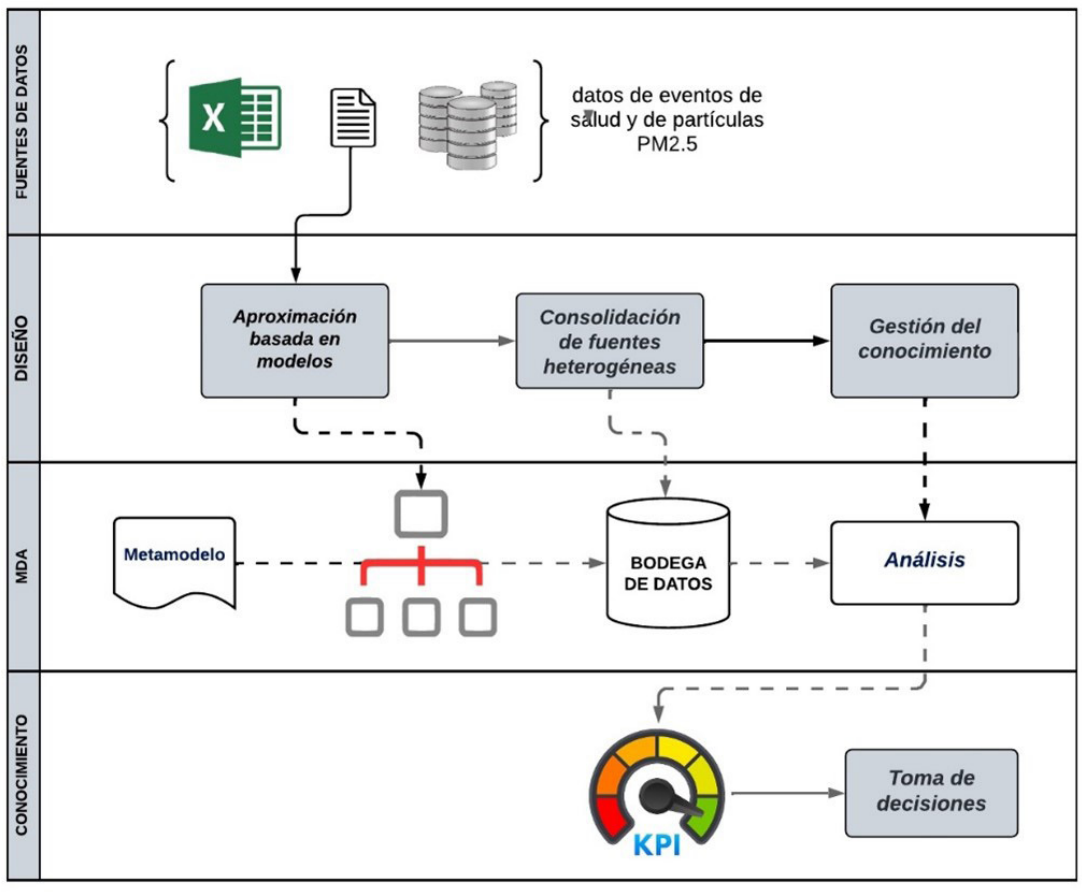
Correspondence
jegirado@elpoli.edu.co
Carrera 48 N° 7-151
Medellín El Poblado.

¿Por qué se realizó?:
Esta investigación se realizó en el marco de un macroproyecto interuniversitario financiado por el Gobierno de Colombia. El objetivo de este proyecto es mejorar las capacidades de innovación e investigación de los departamentos involucrados (Antioquia, Caldas).

¿Cuáles fueron los resultados más relevantes?
El principal resultado de este proyecto es un modelado del dominio de la salud y de las materias particulares. Adicionalmente, propusimos un metamodelo para consolidar los datos de los dominios. También obtenemos una primera aproximación al modelo analítico.

¿Qué aportan estos resultados?
Los resultados de este proyecto pueden utilizarse para generar modelos de big data de diferentes dominios relacionados con la salud pública.

Graphical Abstract



doi: 10.25100/iyc.v26i3.13644

Introducción

El almacenamiento de datos es un proceso que hace referencia al guardado, preservación y organización de información en un formato especifico con capacidad de accesibilidad para una recuperación posterior y un uso por parte de usuarios de los datos (1). En ese orden de ideas, el almacenamiento de datos enmarca las tecnologías, instrumentos, procesos y estándares relacionados con la gestión de información en formato digital. Lo anterior significa, que la información es almacenada en distintas clases de dispositivos de almacenamiento, ya sean discos duros, discos flexibles o discos virtuales en la nube.
También el almacenamiento implica el uso de bases de datos, sistemas de archivos, sistemas gestores de bases de datos, uso de sistemas externos y/o sistemas de estructuras dinámicas de datos. Con ello se facilita el acceso y procesamiento de información sobre distintas fuentes (2).

Para abordar el almacenamiento de datos se emplean diferentes enfoques entre los que se destacan: i) los enfoques basados en ontologías de dominio (3), ii) basados en gestión de procesos de negocio (4), y iii) basado en bodegas de datos (*Datawarehouse*). Estos enfoques tienen una característica en común, la cual es el análisis de los datos previo, así mismo, el uso de estándares apropiados para el modelado de datos.

Uno de estos estándares es el metamodelo CWM *(Common Warehouse Metamodel*), que es un conjunto de especificaciones para la estandarización de la forma que se representan los modelos de bases de datos (esquemas), modelos de transformación de datos, modelos OLAP y de minería de datos, entre otros (5). El principal objetivo de CWM es facilitar el intercambio de metadatos, en un contexto de metamodelos, con una bodega de datos en entornos heterogéneos distribuidos.

Model Driven Architecture (MDA), es una arquitectura que unifica cada paso del ciclo de vida de desarrollo de software, utilizando "*Metamodelos*" para describir las funcionalidades y los requerimientos de desempeño de una aplicación (6). MDA tiene uso en la integración, transformación, traducción y alineación de notaciones, formatos y/o esquemas con características heterogéneas. Como por ejemplo, Saiji, et at. (7), que hacen uso de MDA junto con BPMN (*Business Process Managment Notation*) para transformar de manera semiautomática modelos conceptuales a modelos específicos de una plataforma tipo web.

Algunos trabajos relacionados en esta área son el de Sun, et al. (8) y el de Azzaoui, et al. (9), en donde proponen la implementación de unas soluciones basadas en MDA, para la construcción de grafos de conocimiento a partir de bases de datos relacionales y para la generación de esquemas multidimensionales a partir de bodegas de datos y con un metamodelo. También Belkadi (10), diseñan una solución para la transformación de reglas de negocio descritas en SQL hacia una base de datos NoSQL.

Otro trabajo relacionado es el presentado por Xie et al. (11), donde se propone un mecanismo para minar datos de orígenes heterogéneos, haciendo uso de MDA. También Hanine et al. (12) presentan un novedoso método para la construcción de esquemas conceptuales a partir de bases de datos relacionales, empleando, al igual que el trabajo anterior, un enfoque multidimensional. Por su parte Esbai et al. (13), realizan una aproximación basada en MDA para la generación automática de bodegas de datos a partir de la información de reglas de negocio e indicadores de rendimiento.

Estos trabajos reflejan el potencial que tiene MDA para la gestión, procesamiento, almacenamiento e integración de datos, principalmente con el uso de arquitecturas de bases de datos, específicamente, las bodegas de datos. Estas últimas facilitan el trabajo con MDA, ya que sus procesos y estructuras internas, pueden verse como parte del enfoque basado en modelos. Es por ello, la motivación de poder integrar datos para su almacenamiento, relacionados con material particulado y eventos de salud, aspecto clave para la toma de decisiones en el ciclo de la gestión de la calidad de aire.

doi: 10.25100/iyc.v26i3.13644

Dada la magnitud y complejidad de los datos generados en diferentes áreas científicas, entre ellas, la operación de redes de monitoreo y vigilancia de la calidad de aire, con relación a diferentes contaminantes como, el material particulado (*PM – Por sus siglas en inglés*), contaminante criterio de alto impacto en la salud, la información generada en su medición es robusta, siendo factible analizarla bajo enfoque basado en modelos.

Un área que tiene relación directa con el PM, es la gestión de eventos de salud, ya que precisamente los efectos generados por la presencia de PM, hace que las personas tengan afectaciones. Los eventos en salud son registrados a partir de la información de las asistencias de pacientes a los servicios de salud en las ciudades, ya sean servicios de tipo urgencia o atención básica (14). A partir del análisis de los eventos de salud, es posible diseñar sistemas de vigilancia, que permitan la detección de comportamiento de una comunidad y a partir de ello se puedan definir estrategias para su mejoramiento (15).

Los sistemas de vigilancia en salud se caracterizan por tener un enfoque sistémico para la recopilación, análisis e interpretación de datos de salud, la cual puede ser accedida de manera periódica. El principal objetivo de estos sistemas es la detección y monitoreo de enfermedades y otros eventos importantes para la salud de una población determinada (15). Con ello es posible realizar detección temprana de problemas de salud, evaluación de impactos de las intervenciones de saludo pública y el registro de las evidencias para la toma de decisiones en el campo de la salud.

No obstante, pese a los avances en la gestión y procesamiento de datos relacionados con material particulado y eventos de salud, esots siguen caracterizados por tener una alta complejidad, dada en términos de la diversidad de formatos, la heterogeneidad en las estructuras, los distintos mecanismos de acceso y las diversas arquitecturas de almacenamiento. Por lo anterior, es complejo realizar procesos como el análisis de datos, el procesamiento inteligente de información y la gestión del conocimiento.

Este documento presenta una propuesta basada en MDA para el almacenamiento de datos relacionados con el material particulado PM2.5 y eventos de salud, que vienen siendo generados en la ciudad de Medellín como parte de sendos procesos de investigación que se vienen liderando desde varias instituciones de educación superior de la ciudad junto con la autoridad ambiental y entes del sector salud, a partir del diseño de un metamodelo y utilizando el estándar CWM, que permita su uso en bodegas de datos y el diseño de una arquitectura que soporte la gestión del conocimiento.

Metodología

Para el desarrollo de esta investigación se propone el uso de MDA para el almacenamiento de datos ambientales de salud. El enfoque principal es la participación de expertos en la construcción de artefactos (diagramas conceptuales) que permitieron la identificación del modelo independiente de la plataforma (*Platform Independent Model -PIM-*).

A partir de ello, se propuso un metamodelo, que a su vez permitió la definición de instancias propias de un modelo de definición de la plataforma (*Platform Definition Model -PDM-*) y así finalmente lograr una aproximación al modelo específico de la plataforma (*Platform-Specific Models -PSM-*). La figura 1 resume la metodología de trabajo, en ella se detallan las actividades realizadas y los artefactos generados en la aplicación de la MDA.

doi: 10.25100/iyc.v26i3.13644

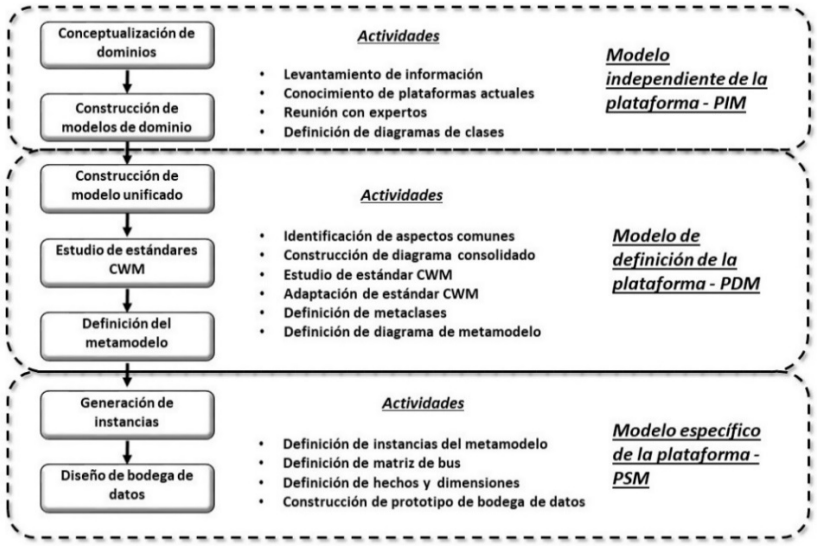


Figura 1. Metodología de trabajo para la aplicación de MDA. Fuente: autor

Como punto de partida se realizó una conceptualización de los dominios, en donde se hizo un levantamiento de información, el reconocimiento de las plataformas de almacenamiento de los datos junto con sus esquemas de representación. Luego, junto con los expertos de cada dominio se logró la construcción de los modelos conceptuales en diagramas de clase UML.

Una vez se construyeron los modelos conceptuales, se procedió a la identificación de los aspectos en común y así obtener un modelo conceptual unificado. Además, se logró una integración conceptual a partir de un enfoque de abstracción de los datos, es decir, la definición de metadatos. Con ello, se definió una extensión para el estándar CWM, a partir de la definición de metaclases y un diagrama del metamodelo.

Una vez se obtuvo el metamodelo, se procedió a la generación de instancias relacionadas con los dominios de calidad de aire y salud pública. Lo anterior se logró con el diseño lógico de la bodega de datos, en donde se incluye la definición de las dimensiones y sus respectivos hechos. Con ello, se pudo identificar elementos claves de unificación de los modelos, como los son la fecha.

Como mecanismo de trabajo para la validación, se empleó una aproximación basada en prototipos. Las conceptualizaciones y decisiones sobre el proyecto se trabajaron sobre diagramas conceptuales, los cuales se iban refinando – a modo de evolución de prototipos - en reuniones periódicas. También cabe señalar, que las validaciones se realizaron con base en la opinión de los expertos y en el contraste con las fuentes de datos.

Modelo de dominio de ambiente - caracterización de material particulado PM2.5

La figura 2 presenta el modelo conceptual de la caracterización química del PM2.5. El PM se define como una mezcla compleja de sustancias orgánicas e inorgánicas, sólidas y líquidas suspendidas en el aire de diámetros aerodinámico entre 50 µm y menores a 2.5 µm (PM2.5) (16), este último de efectos indeseables en la salud de sus habitantes, inhalable por su capacidad de ingresar a alveolos pulmonares, acumularse, traspasar mucosa pulmonar, ingresar al pulmón y en algunos casos transportarse a través del torrente sanguíneo y llegar a otro tipo de órganos. Por este motivo es considerado uno de los contaminantes atmosféricos más dañinos debido a su efecto sobre el deterioro local y regional de la calidad del aire y efectos severos sobre enfermedades respiratorias, cardiovasculares.

doi: 10.25100/iyc.v26i3.13644

Esta afectación está relacionada directamente con su composición química incluyendo elementos y compuestos diversos, mayoritariamente materia orgánica incluyendo los hidrocarburos aromáticos policíclicos, sulfatos, nitratos, amonio, cloruro de sodio, minerales y agua, cuya asociación con un estimado de 7 millones de muertes al año (17), hoy se considera de alta importancia para la consolidación de información robusta y válida sobre este tema en relación con los eventos de salud. El impacto sobre el sistema celular es de tipo biológico, genotóxico o citotóxico.

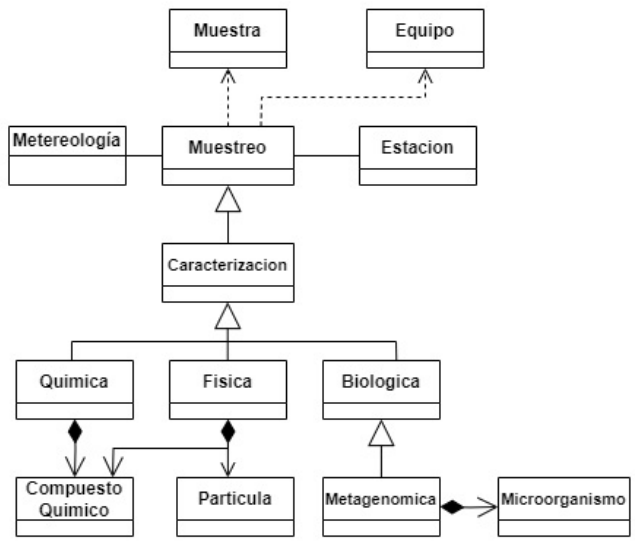


Figura 2. Modelo Conceptual de caracterización química y microbiológica de PM2.5. Fuente: autor

En este diagrama se específica que para la captura de las muestras de PM2.5, se emplean equipos especiales que son ubicados en estaciones de monitoreo. Una vez las muestras son tomadas, son llevadas al laboratorio para realizar su respectiva caracterización. Los tipos de caracterización son: la química, la física y la biológica. Luego se asocian las características propias de la muestra, como su etiqueta y su concentración, y por medio de la fecha de muestreo, se relacionan características meteorológicas que interfieren en el análisis. Una vez recuperadas las muestras y puestas a disposición en laboratorio, se clasifican según su etiqueta y lote, lo cual permite identificar el tipo de caracterización que será realizado para cada una de ellas.

Para la caracterización química, se dispone de información de todos los componentes analizados, determinando en cada una de las muestras su concentración (%) y en microgramo por metro cúbico (µg/m3). Para la caracterización física, se vinculan las partículas que serán motivo de estudio, por ende, todas las partículas que sean reportadas en la caracterización física deberán estar registradas como partículas.

Por su parte, la caracterización microbiológica puede ser de tipo metagenómica, y tiene asociado impactos de tipo genotóxico o citotóxico. Con respecto al análisis genómico, se presenta una situación muy similar a la mencionada en las caracterizaciones química y física, en donde se cuenta con una clase de microorganismos a partir del ADN. En este caso, las principales variables de interés serán determinar la cantidad de microorganismos y de genes en cada muestra. Por último, el análisis genotóxico busca determinar el potencial mutagénico presente en cada lote de muestras.

Actualmente, se cuenta con la posibilidad de obtener la información genómica directamente de las comunidades microbianas en sus hábitats naturales con el fin de inferir perfiles taxonómicos y funcionales de una comunidad microbiana. De esta manera, la metagenómica proporciona la habilidad de estudiar los microorganismos desde el nivel genómico con el fin de entender las relaciones entre los microrganismos, las comunidades y los hábitats en los cuáles ellos viven (18).

doi: 10.25100/iyc.v26i3.13644

Los estudios metagenómicos asociados a la calidad del aire son de gran importancia para la comprensión del impacto potencial de los microrganismos en la salud humana por medio del monitoreo periódico del microbiota asociado al aire. Estos estudios pueden llevar también al descubrimiento de nuevos genes o rutas metabólicas relevantes en aplicaciones industriales, meteorológicas, bioremediación medioambiental y ciclos biogeoquímicos (19).

Algunos estudios han asociado los microorganismos presentes en la atmósfera a diferentes enfermedades humanas, por ejemplo, hongos identificados como causantes de problemas respiratorios (20), síndrome de ganglios linfáticos, enfermedad de Kawasaki (21), al igual que endotoxinas de bacterias del aire han estado también asociadas a problemas de salud (22).

Hasta hace poco tiempo, muchos de los estudios de diversidad microbiana en aire dependían de métodos basados en medios de cultivo, sin embargo, métodos independientes de cultivos usando tecnologías de secuenciamiento de ADN son ampliamente empleados últimamente. Esta tecnología permite tener un panorama más amplio de la diversidad de microorganismos y de esta manera comparar su variación temporal bajo diferentes condiciones meteorológicas.

De esta manera, la metagenómica proporciona la habilidad de estudiar los microorganismos desde el nivel genómico con el fin de entender las relaciones entre los microrganismos, las comunidades y los hábitats en los cuáles ellos viven (18).

Modelo de dominio de eventos de salud

Para MinSalud (23), un evento de salud se relaciona con las "circunstancias que pueden incidir en la situación de salud de un individuo o comunidad. Los eventos de salud se clasifican en condiciones fisiológicas, enfermedades, discapacidades y muertes; factores protectores y factores de riesgo relacionados con condiciones del medio ambiente, consumo y comportamiento; acciones de protección específica, detección temprana y atención de enfermedades y demás factores determinantes asociados."

La figura 3 presenta el modelo conceptual de eventos de salud representado en un diagrama de clases UML, donde se plasman las diferentes etapas de recolección e integración de la información de esta dimensión, como producto de la implementación de un estudio de cohorte en salud ambiental (24), el cual tiene como objetivo investigar la relación entre la exposición al material particulado y la aparición de eventos de salud en una población a lo largo del tiempo (25).

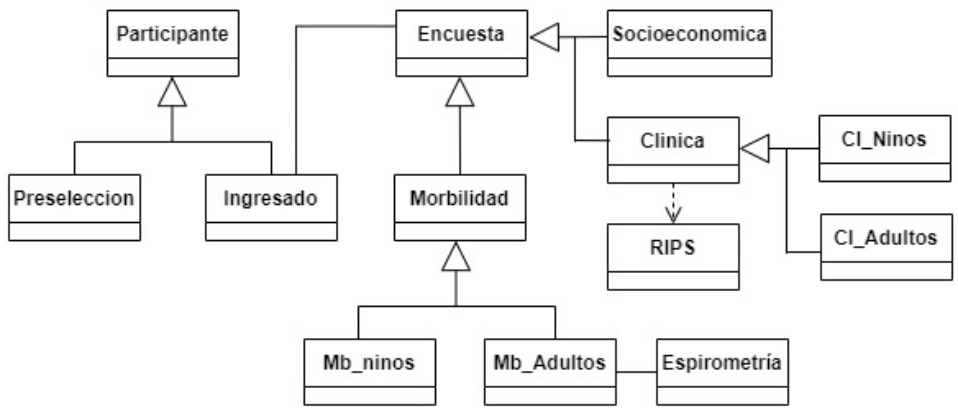


Figura 3. Modelo Conceptual de Eventos de Salud. Fuente: autor

En este estudio se ha conformado una cohorte fija de hombres y mujeres de los grupos etarios de menores de 15 años y mayores a 44 años, los cuales previamente fueron seleccionados, considerando aspectos como: antecedentes clínicos, el tiempo de residencia en la zona de estudio, la aceptación a participar y de brindar información sobre su estado de salud y sus condiciones de

doi: 10.25100/iyc.v26i3.13644

vida, y la autorización para acceder a sus registros de atención en los servicios de salud. La selección de los participantes hace parte de un estudio de cohorte, y sus características no hacen parte del diseño del estudio.

Los participantes ingresaron a la cohorte a partir de un proceso de preselección previa de familias de las zonas de estudio. Las personas que se encuentraban identificadas como ingresados, se les aplicaron diferentes procedimientos de recolección de información: examen físico, encuesta clínica acorde a su grupo etario, encuestas de tipo socioeconómica y encuestas de morbilidad sentida.

También se accedió a los Registros Individuales de Prestación de Servicios en Salud (RIPS), que recopila todas las atenciones de un individuo. Adicionalmente, en un subgrupo de participantes la información en salud se complementó con información de función pulmonar mediante pruebas de Espirométricas (26). Por último, para ampliar el espectro de análisis de los impactos en salud del PM2.5, a una escala molecular, se realizó un análisis de genotoxicidad mediante la prueba Ames, que busca determinar el potencial mutagénico presente en cada lote de muestras de PM 2.5 en momentos de contingencia ambiental.

Aplicación de MDA

MDA permite la creación de modelos altamente abstractos y legibles por máquinas que se desarrollan independientemente de la tecnología de implementación y se almacenan en repositorios estandarizados. Las herramientas pueden acceder a ellos repetidamente y transformarlos automáticamente en esquemas, esqueletos de código, casos de prueba, código de integración y scripts de despliegue para diversas plataformas (27).

La figura 4 presenta la arquitectura de modelado tradicional del OMG, que consiste en una jerarquía de niveles de modelo, cada uno (excepto el superior) caracterizado como “una instancia” del nivel superior. El nivel inferior, también denominado M0, contiene elementos del mundo real correspondientes a los “datos del usuario”, es decir, objetos de datos reales para para los que el software ha sido diseñado para manipular, mientras que el nivel M1 contiene modelos de los objetos del mundo real. El nivel M2 contiene modelos de los modelos en M1, y el nivel M3 contiene modelos de los modelos en M2 (28).

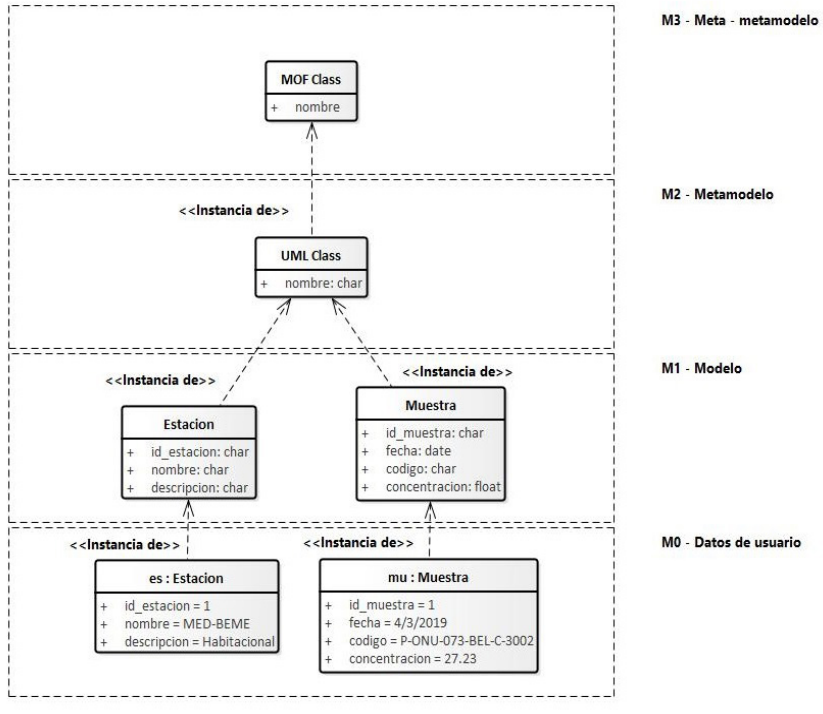


Figura 4. Arquitectura de 4 niveles propuesta por OMG. Fuente: autor

doi: 10.25100/iyc.v26i3.13644

Para la definición del metamodelo orientado a metadatos, se hace uso del estándar Common Warehouse Metamodel (CWM) alineado a lo propuesto por Meta-Object Facility (MOF). Este último proporciona el estándar de modelado reside en la capa M3. Específicamente para el problema del almacenamiento de datos de la caracterización del PM2.5 y los eventos de salud, las capas de la arquitectura MDA son los siguientes:

Capa M0: esta capa tiene como entrada inicial los datos estructurados de cada uno de los dominios representados a su vez como bases de datos, hojas de cálculo, archivos planos y archivos físicos. Esta capa es la de más baja jerarquía de los modelos y representa el nivel más concreto y detallado del metamodelo.

Para el dominio de PM2.5 se tiene una representación de datos por medio de archivos planos en formato .cvs y .txt. Los datos de PM2.5 y la caracterización de sus muestras, comprende el procesamiento de los archivos fuentes para que sean interpretados por los modelos de dominio. Para el caso de los datos relacionados con el dominio de eventos de salud, las encuestas reposan en bases de datos relacionales ligadas con aplicaciones en ejecución. Así mismo, se cuenta con hojas de cálculo para la automatización en macros.

Capa M1: esta capa comprende la instancia del metamodelo, es decir, los modelos de dominio definidos (de caracterización y de eventos de salud), junto con sus respectivas reglas de negocio con el cual es posible razonar de manera integrada. Principalmente, se obtuvo un modelo conceptual a partir de los datos presentes en la capa M0. En dichos modelos se logró reflejar la lógica y la funcionalidad básica alrededor de los datos.

En esta capa M1 se tiene una visión independiente de la plataforma, ya que en ella se describe la lógica y el comportamiento del sistema, sin llegar a detalles específicos. Una ventaja de ello es la capacidad de generar esquemas que puedan ser aprovechados por otros dominios, que por ahora no se han tenido en cuenta, como lo es la información meteorológica o la satelital.

Para modelar la capa M1 se tuvo en cuenta una notación específica, esta fue la de diagramas de clases UML, con los cuales se logra una formalidad aceptada para la comunidad. A esto se suma que, UML comprende extensiones para representación de meta clases y meta-metaclases, lo que facilita la definición de esquemas para la adaptación a modelos específicos de la plataforma.

Capa M2: en esta capa está presente el metamodelo propuesto. Principalmente, a partir de esta capa es posible generar instancias en donde se visualiza de una manera integrada aspectos comunes de los modelos de la capa M1. Es decir, a partir de la capa M2, se puede hacer una representación específica de la plataforma, en este caso para bodegas de datos.

Los modelos generados en la capa M2 son una abstracción de la capa M1, donde se refleja la plataforma específica de ejecución de los datos, ya sea un esquema de datos o una tecnología específica. En concreto, para la investigación, se define el metamodelo con base en lo propuesto por el estándar CWM. Es así como las instancias de la capa M2, están estructuradas de manera tal que pueden ser fácilmente llevadas a bodegas de datos.

A partir de esta capa es posible obtener un detalle de la implementación de sus instancias, ya que se espera, que lo modelos generados en la capa M2, sirvan como base para la generación de código fuente que pueda ser implementado en herramientas de software. Lo anterior significa que se fortalece la capacidad de reescribir a lógica de ejecución de acuerdo con la plataforma definida.

Capa M3: no aplica para los alcances de esta investigación.

doi: 10.25100/iyc.v26i3.13644

Resultados

Metamodelo propuesto (Capa M2)

El metamodelo propuesto tiene como base las metaclases esquemas, tablas y columnas. Con estas metaclases se pueden generar instancias del metamodelo apropiadas para la integración de datos de los modelos de dominio de caracterización y eventos de salud. Luego se definió el metamodelo que complementa el estándar CWM. La figura 5, presenta en notación UML el metamodelo propuesto. Se define como paquete central al estándar CWM, el cual se complementa por un modelo de gestión de esquemas de integración y un otro para la gestión de los datos generados en los dominios.

Respecto a los modelos de dominio, la introducción del patrón de diseño Factory Method en el metamodelo, presenta de manera eficiente, cómo gestionar la integración de la diferente información de los dominios, específicamente en las áreas de ambiente y salud. Este patrón fue aplicado con base en la experiencia del equipo de trabajo. El uso del patrón Factory Method permite de manera efectiva la creación de objetos, facilitando la extensibilidad del metamodelo al incorporar nuevos esquemas de dominio de manera sencilla y organizada.

La adopción del patrón Factory Method brinda flexibilidad y escalabilidad al modelo, donde la creación de instancias de metaclases, relacionadas con los dominios de ambiente y salud, se gestiona de manera centralizada, proporcionando un estándar coherente para la incorporación de futuros esquemas de dominio.

La adopción del estándar CWM en el diseño del metamodelo, facilita la representación coherente y estructurada de los metadatos, que describen tanto bases de datos relacionales como la complejidad de una bodega de datos. Es así como un esquema en “estrella” puede representarse como una base de datos relacional. Normalmente, las bodegas de datos son bases de datos relacionales, las cuales se procesan en un sistema de gestión de bases de datos relacionales (29).

El diseño propuesto proporciona un marco unificado para la descripción de esquemas de dominio diversos, permitiendo la integración fluida de datos correspondientes a los dominios de ambiente y salud, así como la incorporación de otros posibles esquemas de dominio en el futuro. La estandarización ofrecida por CWM simplifica la relación de los metadatos, permitiendo una gestión eficiente de la información en la bodega de datos.

doi: 10.25100/iyc.v26i3.13644

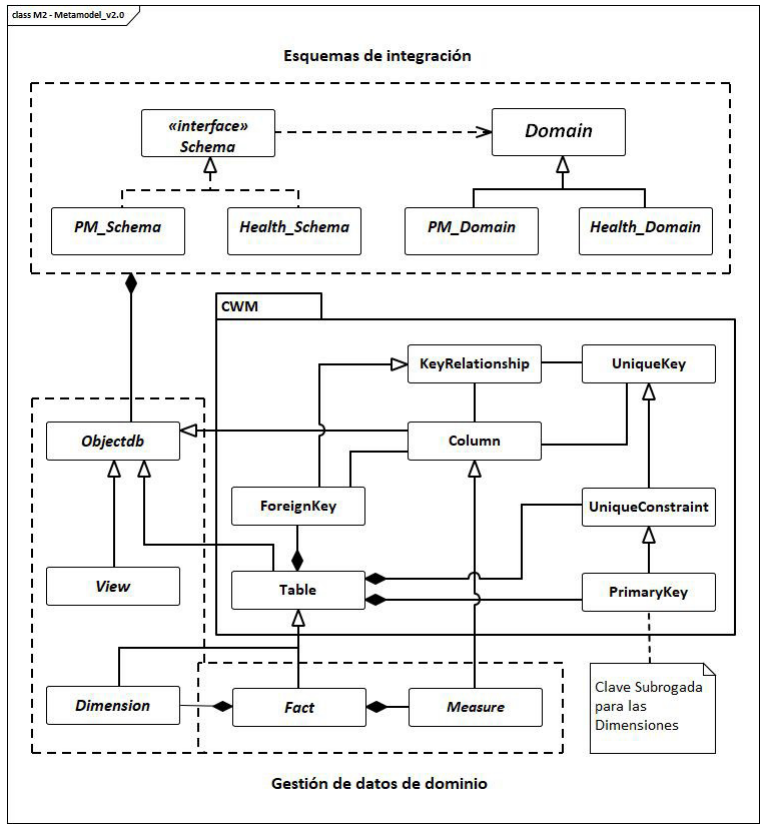


Figura 5. Metamodelo propuesto. Fuente: autor

Este enfoque basado en metadatos presenta ventajas al garantizar la coherencia en la relación de los datos, independientemente de la complejidad y diversidad de los esquemas de dominio, definiendo de manera uniforme objetos de bases de datos clave como tablas, vistas o columnas, además de las restricciones entre estas, lo que facilita la comprensión y el mantenimiento del metamodelo. Así mismo, la flexibilidad del estándar CWM brinda una base sólida y escalable, permitiendo al metamodelo adaptarse a nuevas exigencias y esquemas de dominio sin comprometer la integridad y la consistencia de la bodega de datos. En este sentido, las columnas que actúan como clave primaria en las tablas de dimensiones son claves subrogadas generadas artificialmente para identificar cada registro de manera única en la tabla de dimensiones (30).

Por último, la implementación de un enfoque multidimensional como parte del diseño del metamodelo de la bodega de datos representa de manera integral y estructurada, la gestión de los datos generados en los dominios al utilizar dimensiones, hechos y medidas, permitiendo una comprensión más clara de las relaciones entre los datos almacenados. De este modo, es posible clasificar y organizar los datos en categorías relevantes que describan los dominios, así como los eventos o fenómenos que se presentan en estos.

Dentro del diseño también se introduce una metaclase cuyas instancias hacen referencia a las variables que permiten la cuantificación y análisis de datos, facilitando el almacenamiento de información heterogénea de manera coherente, además de enriquecer el metamodelo al posibilitar que los datos relacionados con los dominios de ambiente y salud se representen de una manera más completa y comprensible para la toma de decisiones.

Implementación y validación

A partir del metamodelo propuesto se presenta el diseño preliminar de un prototipo de una bodega de datos que permita validar las instancias que se generan desde le metamodelo. Inicialmente se presenta el diseño de la matriz de bus, para posteriormente definir modelo estrella que la representa.

doi: 10.25100/iyc.v26i3.13644

Matriz de bus propuesta

En el ámbito de la investigación sobre el impacto de la contaminación del aire en la salud humana la construcción de una matriz de bus se constituye como un recurso fundamental diseñado con la finalidad primordial de facilitar la integración y el análisis de datos multidimensionales, permitiendo así una comprensión más profunda de la relación entre la exposición a contaminantes atmosféricos y los efectos en la salud. En una matriz de bus la granularidad se refiere al nivel de detalle en el que se registran los hechos o medidas, mientras que la jerarquía se relaciona con la organización estructurada de las dimensiones a diferentes niveles de agregación o detalle (31).

En la tabla 1, se presenta datos provenientes de diversas fuentes, incluyendo mediciones de concentraciones de contaminantes (expresadas en microgramos por metro cúbico), características meteorológicas (temperatura, velocidad y dirección del viento, pluviosidad y radiación), y atributos contextuales (tiempo, ubicación geográfica de estaciones de monitoreo, detalles sobre participantes en el estudio, información sobre equipos de monitoreo y elementos químicos específicos presentes en las muestras de material particulado).

Tabla 1. Matriz de bus para el diseño de la bodega de datos

Dimensiones	**Hechos**		
	Concentración	**Defunciones**	**Atenciones**
Tiempo	●	●	●
Geografía	●	●	●
Estación	●		
Equipo	●		
Elemento	●		
Participante		●	●
Diagnóstico		●	●
Servicio			●
Medicamento			●
Procedimiento			●

Las celdas de esta matriz almacenan datos específicos que representan la concentración de partículas PM2.5 en una ubicación geográfica y en un momento determinado, así como la cantidad de atenciones médicas y defunciones relacionadas con la exposición a esa concentración. Las dimensiones (Tiempo y Geografía) permiten organizar y contextualizar estos datos, lo que facilita su análisis multidimensional.

Para la validación del metamodelo se recopilaron datos de concentración del contaminante P.M.2,5 con frecuencia de muestreo de cada 3 días. Se incluyen las dimensiones de tiempo, la geografía y estación de monitoreo. Cada una de estas dimensiones tiene unas jerarquías que permiten analizar datos a diferentes niveles de detalle. Para ello se atendió el problema de las distintas escalas de tiempo, a modo de ejemplo, la dimensión “Tiempo” tiene una jerarquía que abarca desde el nivel más alto (año) hasta niveles más detallados (mes, día) para permitir análisis temporales a diferentes escalas. La dimensión “Geografía” incluye jerarquías que van desde el nivel más alto (Municipio) hasta niveles más detallados (Comuna, Barrio, Manzana, Coordenadas del punto de monitoreo).

doi: 10.25100/iyc.v26i3.13644

Las columnas de esta matriz representan los hechos a analizar según las dimensiones que los describen. Para los tres hechos se requerirá realizar un análisis desde las dimensiones de Tiempo y Geografía. Para describir el hecho del nivel de concentración del contaminante P.M.2,5 se necesitará incluir además aspectos como la Estación y el Equipo de monitoreo en donde se recolectó la muestra, y adicionalmente, los elementos que componen cada muestra.

Por su parte, los hechos relacionados con la morbimortalidad (tanto atenciones como defunciones) requieren adicionar dimensiones que presenten datos contextuales del participante de la cohorte y su diagnóstico determinado, mientras que aquellos consignados en los registros médicos obtenidos durante el proceso de atención en los diferentes servicios de salud y los medicamentos y procedimientos suministrados serán esenciales únicamente en el hecho de atenciones.

Prototipo de esquema en estrella para la bodega de datos

La construcción de la bodega de datos para la integración de información de salud y ambiente como una instancia del metamodelo de datos, plantea desafíos notables en términos de heterogeneidad de fuentes de datos, calidad, privacidad y escalabilidad. Para ello se definió un esquema en estrella a partir de la matriz de bus propuesta. En la figura 6 se presenta el esquema en estrella definido, cuyo hecho (fact) principal es la Morbilidad, del cual dependen las dimensiones de Geografía, Participante y Tiempo.

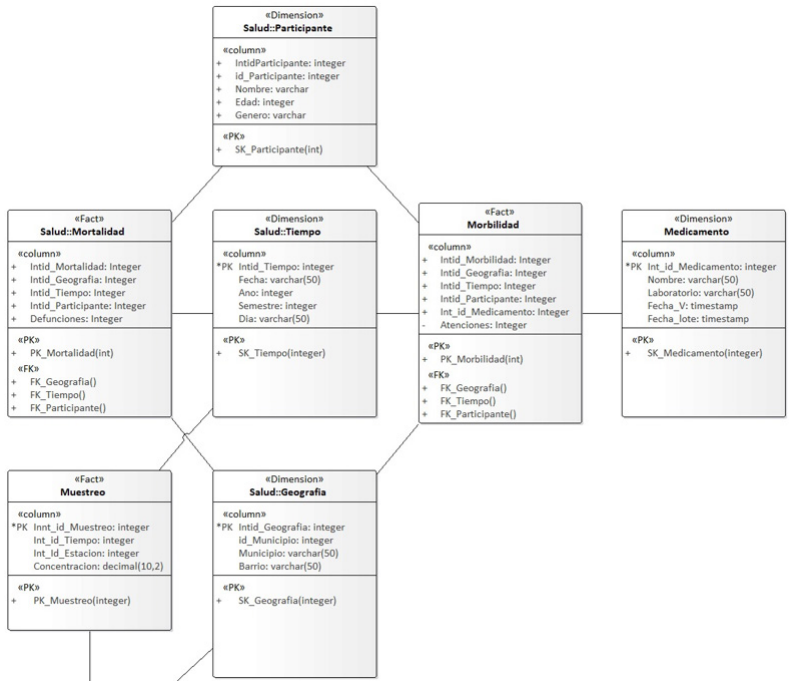


Figura 6. Esquema en estrella para el dominio de Salud

Discusión y análisis

La gestión de datos faltantes y la corrección de errores de registro son procedimientos esenciales para evitar sesgos en el análisis. Además, la disponibilidad de datos presenta desafíos en términos de cobertura geográfica y continuidad temporal, lo que requiere estrategias de recopilación y tratamiento de datos efectivas. La complejidad de las relaciones entre la exposición a contaminantes y los efectos en la salud demanda enfoques analíticos avanzados para capturar relaciones no lineales y controlar factores de confusión. Del mismo modo, la privacidad y seguridad de los datos personales de los participantes son consideraciones fundamentales.

doi: 10.25100/iyc.v26i3.13644

A lo anterior se suma que, la escalabilidad de la infraestructura de datos y el mantenimiento constante de registros son críticos a medida que se acumulan más datos con el tiempo. Por último, la comunicación efectiva de los resultados a audiencias técnicas y no técnicas es esencial para traducir los hallazgos y decisiones informadas en el ámbito de la salud pública y la gestión de la calidad del aire. La principal dificultad del manejo de los datos en el nivel M0 del metamodelo, es la constante modificación y riesgos ante posible corrupción de estos. Lo anterior se refleja cambios de fuentes y extensiones de archivos. Del mismo modo existe una dependencia respecto a los procesos de mantenimiento y privacidad. Estas dificultades pueden afectar las otras capas, por ello la importancia de la parametrización.

En el caso de la información ambiental también se cuenta con distintos tipos de datos de acuerdo con la caracterización química, los niveles de cada elemento, compuesto específico y otros potenciales análisis en relación con la agrupación de estos elementos y compuestos en perfiles más robustos de acuerdo con fuentes específicas de contaminación.

Se debe tener presente el componente temporal en este proceso de integración, un diseño longitudinal implica la integración de datos ambientales y de salud en diferentes marcos temporales, es decir, datos que se toman en varios momentos. De salud se cuenta con datos desde el 2008 a la fecha, de distintas fuentes, que se han tomado en varios momentos o en un solo momento, según los objetivos del estudio. Así mismo, el metamodelo propuesto debe poder integrar diseños anidados a los análisis, por ejemplo, análisis transversales a la cohorte, casos y controles anidados, estudios de panel o diseño de cohorte-cohorte. Esto le permitirá al metamodelo ampliar las posibilidades de integración de datos.

Por otro lado, las ventajas inherentes a esta construcción son sustanciales. La sinergia entre datos de salud y ambiente puede propiciar una comprensión más holística de las relaciones causales entre exposición y efectos en la salud humana. Esto, a su vez, puede informar de manera más precisa la formulación de políticas públicas y estrategias de salud, y ofrecer la oportunidad de investigaciones científicas más sofisticadas.

La detección temprana de tendencias adversas en la salud pública relacionadas con la calidad del aire puede brindar una ventaja preventiva y promover una toma de decisiones más informada, mientras que la transparencia en la disponibilidad de datos y modelos puede fomentar la confianza pública y científica en los esfuerzos de mitigación ambiental y de salud.

Conclusiones

El trabajo presentado destaca la posibilidad de consolidar, mediante un metamodelo datos relacionados con material particulado PM2.5 y eventos de salud pública. A partir de ello, es posible almacenar datos de una manera coherente, desde el punto de vista conceptual, con el fin de facilitar la toma de decisiones y promover programas de protección de la salud.

Los resultados derivados de este enfoque analítico pueden ser utilizados como base sólida para la toma de decisiones en políticas de salud pública y estrategias de gestión de la calidad del aire. La información obtenida puede contribuir al diseño de medidas de control de la contaminación más efectivas y a la formulación de recomendaciones fundamentadas para la preservación de la salud de la población expuesta.

La adopción de MDA y en especial CWM ha evidenciado ser un enfoque de diseño arquitectónico efectivo para la representación de un metamodelo donde se presentan los metadatos que describen una bodega de datos, sin embargo, la integración de los principios de DDD (Domain Driven Design)

doi: 10.25100/iyc.v26i3.13644

podría ser considerada para mejorar la identificación y delimitación de los dominios específicos dentro de la bodega de datos. DDD brinda un enfoque de diseño sólido para comprender el negocio, permitiendo establecer con claridad los límites de los dominios para una representación coherente de los datos.

Así mismo, marcos de trabajo como DIF (Data Integration Framework) o MIS (Metadata Integration Strategy), representan áreas de interés para investigaciones futuras, con el objetivo de plantear una integración eficiente de los datos, teniendo en cuenta la variedad de fuentes de datos que se pueden almacenar en una bodega de datos, además de incorporar estrategias efectivas para abordar los metadatos de manera eficiente.

Referencias

(1) Elmasri, R, Navathe S, Castillo V, Pérez G, Espiga, B. Fundamentos de sistemas de bases de datos. Pearson educación; 2007.

(2) Jarke, M, Lenzerini, M, Vassiliou, Y, Vassiliadis, P. (2002). Fundamentals of data warehouses. Springer Science & Business Media.

(3) Olivé A. A universal ontology-based approach to data integration. Enterprise Modelling and Information Systems Architectures (EMISAJ), 13, 110-119; 2018

(4) Dumas M, La Rosa M, Mendling J, Reijers, H. Fundamentals of business process management. Springer; 2013.

(5) Poole, J., Chang, D., Tolbert, D., & Mellor, D. (2002). Common warehouse metamodel. John Wiley & Sons; 2002.

(6) Object Management Group Model Driven Architecture (MDA). OMG MDA Guide rev. 2.0; 2014.

(7) Sajji A, Rhazali Y, Hadi Y. A methodology for transforming BPMN to IFML into MDA; Bulletin of Electrical Engineering and Informatics, 2022; 11(5), 2773-2782.

(8) Sun S, Meng F, Chu D. A model driven approach to constructing knowledge graph from relational database. In Journal of Physics: Conference Series (Vol. 1584, No. 1, p. 012073). IOP Publishing; 2020.

(9) Azzaoui A, Rabhi O, Mani A. A model driven architecture approach to generate multidimensional schemas of data warehouses; 2019.

(10) Belkadi F, Esbai R. A Model-Driven Engineering: From Relational Database to Document-oriented Database in Big Data Context. In ICSOFT (pp. 653-659); 2021.

(11) Xie J, Xu F, Li Z, Li X. Data Mining Method under Model-Driven Architecture (MDA). Security and Communication Networks; 2022.

(12) Hanine M, Lachgar M, Elmahfoudi S, Boutkhoum O. MDA Approach for Designing and Developing Data Warehouses: A Systematic Review & Proposal. International Journal of Online & Biomedical Engineering; 2021; 17(10).

(13) Esbai R, Hakkou R, Habri A. Modeling and automatic generation of data warehouse using model-driven transformation in business intelligence process. Indonesian Journal of Electrical Engineering and Computer Science Vol. 30, No. 3, June 2023, pp. 1866~1874 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v30.i3.pp1866-1874

(14) Peláez O, y otros.as Bermejo, P. Brotes, epidemias, eventos y otros términos epidemiológicos de uso cotidiano. Revista Cubana de Salud Pública, 46, e2358; 2020.

(15) Mercuriali, L, Oliveras L, Gómez A, Marí, M, Montalvo T, Villalbí J. Un sistema de vigilancia de salud pública para el cambio climático en las ciudades. Gaceta Sanitaria, 36, 283-286; 2022.

(16) United States Environmental Protection Agency. Particulate matter (PM) basics; 2017.

(17) Novillo-Ortiz D, D'Agostino M, Becerra-Posada F. El rol de la OPS/OMS en el desarrollo de capacidad en

doi: 10.25100/iyc.v26i3.13644

eSalud en las Américas: análisis del período 2011-2015. Revista Panamericana de Salud Pública; 2016; 40, 85-89.

(18) Wooley J, Godzik A, Friedberg I. A Primer on Metagenomics. PLoS Comput Biol 6(2): e1000667. https://doi.org/10.1371/journal.pcbi.1000667; 2010.

(19) Behzad H, Gojobori T, Mineta K. Challenges and Opportunities of Airborne Metagenomics. Genome Biol Evol ;7:1216–1226. doi: 10.1093/gbe/evv064; 2015

(20) Grinn-Gofroń A, Strzelczak A. Changes in concentration of Alternaria and Cladosporium spores during summer storms. Int J Biometeorol. Sep; 57(5):759-68; 2013

(21) Rodó X, Curcoll R, Robinson M, Ballester, J, Burns, J, Cayan R., ... Morguí, J. Tropospheric winds from northeastern China carry the etiologic agent of Kawasaki disease from its source to Japan. Proceedings of the National Academy of Sciences, 111(22), 7952-7957; 2014.

(22) Mueller-Anneling L, Avol E, Peters JM, Thorne PS. Ambient endotoxin concentrations in PM10 from Southern California. Environ Health Perspect. Apr; 112(5):583-8; 2004.

(23) Ministerio de Salud, S. D. S., & Inspección, S. (2006). DECRETO 3518 DE 2006 (OCTUBRE 09).

(24) Lazcano-Ponce E, Fernández E, Salazar-Martínez E, Hernández-Avila, M. Estudios de cohorte. Metodología, sesgos y aplicación. Salud pública de México, 42, 230-241; 2000.

(25) World Health Organization. WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: executive summary; 2021.

(26) Taylor J, Shrubsole C, Symonds P, Mackenzie I, Davies, M. Application of an indoor air pollution metamodel to a spatially-distributed housing stock. Science of the Total Environment, 667, 390-399; 2019.

(27) Kleppe A, Warmer J, Bast W. MDA explained: the model driven architecture: practice and promise. Addison-Wesley Professional; 2003.

(28) Atkinson C, Kühne T. Model-driven development: A metamodeling foundation. IEEE Software, 20(5), 36–41. https://doi.org/10.1109/MS.2003.1231149; 2003.

(29) Imran, S., Mahmood, T., Qamar, A. M., Siddiqui, A. J., Ahmed, I., & Shariq, N. (2024). NODW Framework for Data Warehousing-A NoSQL Big Data Perspective. Authorea Preprints.

(30) Wijaya, W., & Wiratama, J. (2024). The Implementation of Data Warehouse and Star Schema for Optimizing Property Business Decision Making. G-Tech: Journal Teknologi Terapan, 8(2), 1242-1250.

(31) Kimball R, Ross M. The data warehouse toolkit: the definitive guide to dimensional modeling. John Wiley & Sons; 2013