Paradigmas para el diseño multidimensional de almacenes de datos: un mapeo sistemático

Pablo Enrique Espinoza Navarrete
Ingeniería informática
Universidad de La Frontera
Temuco, Chile
p.espinoza04@ufromail.cl

Ania Lorena Cravero Leal
Departamento de Ingeniería de Sistemas
Universidad de La Frontera
Temuco, Chile
ania.cravero@ufrontera.cl

***Resumen* – Un almacén de datos es una colección de datos orientada a un dominio, integrada, no volátil y variante en el tiempo para ayudar a la toma de decisiones de una organización. Existen propuestas metodológicas para el diseño multidimensional de un almacén de datos basadas en tres paradigmas, estos son el enfoque impulsado por la oferta, impulsado por la demanda y el hibrido. Evaluamos las propuestas existentes con el fin de identificar sus principales características, contribuciones, ámbito de desarrollo y aplicación. Como metodología hemos llevado a cabo el proceso de mapeo sistemático, un tipo de estudio secundario diseñado específicamente para abordar este tipo de objetivo. La principal conclusión es que la gran mayoría de las propuestas son del ámbito académico, muy pocas en la industria y ninguna en base a experimentos.**

I. Introducción

Los almacenes de datos (AD) son diseñados para respaldar la toma de decisiones en la organización [1, 2]; Estos sistemas deben homogenizar e integrar los datos de diferentes áreas de una organización en un gran repositorio de datos, con el objetivo tomar ventaja de una representación única y detallada, permitiendo así la extracción del conocimiento que es relevante para la toma de decisiones.

Sin embargo la construcción de un AD es todavía un gran reto debido a que es una tarea muy compleja interrelacionar componentes con diferentes funciones, diferentes diseños y diferentes tecnologías [3]. Una de las principales dificultades al momento de desarrollar un AD es que aún no hay un modelo estándar definido para ello, sin embargo es ampliamente asumido que el diseño de un AD debe seguir un paradigma multidimensional [3, 4]. Así diferentes aproximaciones y métodos han sido propuestos para diseñar cada capa de un AD multidimensional.

Según Winter y Strauch [5] las aproximaciones o metodologías para el modelado multidimensional pueden ser clasificadas de acuerdo a la forma cómo se obtienen los requerimientos de un AD: Estos son, los enfoques impulsados por la oferta, impulsados por la demanda y el hibrido que es una mezcla de las dos anteriores.

Dada la importancia de los AD hoy en día, en este artículo se brinda un estudio comparativo de las metodologías para el diseño multidimensional de AD a través de un mapeo sistemático de estudios. El mapeo entregará datos como: características, contribuciones, áreas de desarrollo y aplicación. Es con esta motivación que el estudio actual ha surgido de nuestro trabajo para recopilar, mapear y resumir los estudios primarios.

Un mapeo sistemático de estudios es una metodología que se utiliza con frecuencia en la investigación médica, pero que ha sido adecuado para el uso en temas del área informática. El objetivo principal de un mapeo sistemático de estudios es proporcionar una visión general de un área de investigación, determinar la cantidad y tipo de investigación y los resultados disponibles[6]. Requiere menos esfuerzo, que una revisión sistemática, ya que proporciona una visión de más alto nivel de granularidad con el objetivo de identificar áreas donde la investigación es escasa y donde es abundante.

Este trabajo está organizado de la siguiente manera: en la sección 2 se presentan los conceptos básicos y una breve descripción de los paradigmas para el diseño multidimensional de un AD. En la sección 3 se describe el proceso de mapeo sistemático. En la sección 4, se describen los resultados. En la sección 5 se presentan las conclusiones.

II. Conceptos Básicos

La definición clásica de AD fue acuñada por Immon [7] como una colección de datos históricos, orientados por temas, no volátiles, integrados, diseñados para apoyar el proceso de toma de decisiones de una organización. Desde el punto de vista funcional, el proceso de un AD lo componen 3 fases: extracción de datos desde distintas Fuentes de datos, transformación y carga de datos de manera consistente en el AD, el acceso de datos integrados de una forma eficiente y flexible [8]. Las dos primeras etapas forman parte del proceso conocido como Extraction-Transformation-Load (ETL).

La principal contribución de un AD es su capacidad de convertir los datos en información estratégica, apoyando la toma de decisiones en los niveles más altos de una organización. Esto se logra a través de la herramienta OLAP [9] permitiéndoles a los usuarios finales analizar, explorar, navegar por diferentes niveles de detalle.

Los datos se organizan en una forma multidimensional, conocido como esquema estrella, donde la información se clasifica de acuerdo a hechos y dimensiones permitiendo así una manera más rápida y flexible de acceder a los datos a través de consultas con herramientas OLAP [10]. Los hechos son los datos numéricos o un foco de interés (actividad específica) para el proceso de toma de decisiones. Las

dimensiones son las perspectivas individuales de los datos que determina la granularidad (datos a nivel de detalle) que se adopten para la representación de un hecho. Una medida es una propiedad numérica de un hecho y describe uno de sus aspectos cuantitativos de interés para el análisis.

Paradigmas para el diseño del almacén de datos

A continuación se describen las características principales de los distintos paradigmas para el diseño de ADs.

Paradigma impulsado por la oferta

Estos enfoques impulsados por la oferta (también conocido como impulsado por los datos), inician el proceso de modelado del AD desde un análisis detallado de las fuentes de datos para determinar los elementos (como hechos dimensiones), más relevantes para el proceso de toma de decisiones. Generalmente en este enfoque la información almacenada en los hechos representan medidas para los procesos de negocio y busca responder preguntas como ¿Cuál es el producto que más se vende?, ¿Qué medicamentos son los más recetados?, ¿Cuántos pacientes son tratados?, etc. Las dimensiones representan el marco para el análisis de estas medidas (tienda, región, tiempo) [11].

Paradigma impulsado por la demanda

Estos enfoques también son conocidos como impulsados por requisitos o dirigidos a objetivos, empiezan en la determinación de las necesidades de los usuarios, para luego crear un diseño multidimensional del AD de acuerdo a los objetivos seleccionados. Este enfoque proporciona apoyo a situaciones específicas de la organización así como también a introducir cambios en los procesos de negocios. Además permite responder preguntas como si es posible cumplir un objetivo [12].

Paradigma Hibrido

Estos enfoques proponen combinar ambos paradigmas con el objetivo de diseñar el AD desde las fuentes de datos, pero teniendo en cuenta las necesidades del usuario final. La principal característica y diferencia en relación a los dos enfoques anteriores es que este tipo de enfoque tiene la posibilidad de intercalar los enfoques impulsados por la oferta y demanda para aplicarlos en cada etapa del desarrollo del AD, beneficiándose de la información recopilada a lo largo del proceso [12].

III. Proceso del mapeo sistemático

El objetivo principal del mapeo sistemático de estudios llevado a cabo es obtener una visión general de la investigación sobre los paradigmas para el diseño multidimensional de los ADs. Además no solo se pretende identificar las principales aproximaciones en esta área, sino también sus puntos fuertes y debilidades y, por supuesto, el trabajo futuro que puede llevarse a cabo para solventar posibles debilidades. El objetivo general es definido en cuatro preguntas de investigación (PI):

(PI1) ¿Cuál es el paradigma más utilizado en las investigaciones seleccionadas y como ha cambiado la tendencia a lo largo del tiempo? Esto nos permite saber cuál es la tendencia que presenta está área, cual enfoque está vigente y cuál no.

(PI2) ¿Cuál de estos ámbitos (academia, industria) es el más usual al momento de aplicar la investigación? Permite dilucidar el ámbito de preferencia en el cual es aplicada la investigación para diseñar un AD.

(PI3) ¿Cuál es la contribución de los trabajos de investigación al área? Permite identificar el aporte de los trabajos al diseño multidimensional de los ADs.

(PI4) ¿Cuál etapa del diseño de un AD es la más investigada? Permite dilucidar qué etapa es la que tiene mayor investigación y en la que se enfocan la mayoría de los investigadores.

A continuación se presentan las etapas del mapeo sistemático de estudios:

3.1 definición del alcance, selección de la estrategia y criterios de selección.

El alcance de este estudio fue el siguiente: *Población:* Conjunto de artículos que describen los estudios sobre el diseño multidimensional de un AD en la academia, industria. *Intervención:* Cualquier estudio con métodos, metodologías , herramientas, etc. Basadas en los paradigmas para el diseño del AD. *Diseño del estudio:* Experimentos, caso de estudios, relatos de experiencia, la investigación-acción. *Resultado:* cantidad y tipo de evidencia relativas al diseño multidimensional del AD.

La cadena de búsqueda utilizada como base para la obtención de los trabajos relevantes ha sido: ("*data warehouse"* OR *data warehousing*) AND *"multidimensional design" AND (approach* OR *methodology).*Algunos de los términos fueron desglosados en expresiones booleanas de tipo OR y AND, formada por los sinónimos como por ejemplo: *data warehouse, data warehousing*. Con respecto al tiempo, la búsqueda se ha centrado en los años 1998-2013. Esta elección está motivada debido a que a partir del año 1998 varios investigadores se adentraron en este tema, teniendo como base las investigaciones de los "padres" de los almacenes de datos Bill Immon[7] y Ralph Kimball[3].

Por otro lado la selección de las fuentes de datos todas fueron digitales. Se seleccionaron estas fuentes, ya que incluyen motores de búsquedas y los artículos que ofrecen son de calidad, además son accesibles vía web. Las fuentes donde se aplicó la búsqueda fueron *Google Scholar, IEEE y Springer*.

La ejecución de la búsqueda en estos motores específicos arrojó 288 resultados (*Google Scholar*), 4 (*IEEE*) y 1915 (*Springer*) respectivamente. Para comprobar que los buscadores ordenan por relevancia se revisaron los primeros 300 resultados, en el caso de Springer esto permitió comprobar que los resultados encontrados después de esta cifra (300) no son relevantes y arrojan publicaciones que no se

acercan a el objetivo de nuestra búsqueda. Dado que los tres buscadores ordenan por relevancia de la publicación, se decidió analizar los 288 resultados más relevantes de cada buscador, lo que nos da un total de 580 publicaciones analizadas. En la tabla 1 se presenta un resumen de estos resultados.

Buscador	**Google Scholar**	**IEEE**	**Springer**	**Total**
Resultado de la búsqueda	288	4	1915	2207
Trabajos analizados	288	4	288	580
Trabajos candidatos	54	1	38	93
Trabajos relevantes	24	1	9	34
Concidencias con *Google Scholar*	-	1	8	9
Total trabajos relevantes	24	0	1	25

Tabla 1. Resultados de la búsqueda antes y después de eliminar duplicados.

La selección de los estudios se ha formulado basada en los siguientes criterios de inclusión/exclusión:

Inclusión*:* libros, documentos, artículos, tesis, trabajos de investigación, publicaciones de revistas y congresos que describan el diseño multidimensional de un almacén de datos y que contengan aproximaciones, metodologías y herramientas en cualquiera de las siguientes etapas para el diseño multidimensional de un almacén de datos: conceptual, lógica y física.

Exclusión*:*

1. Trabajos que tratan sobre los almacenes de datos, pero que no están relacionados con el diseño multidimensional de éstos. Ej. Experiencias de uso de un AD en la industria o en la academia, análisis de datos con AD, Business Inteligence, Data mining, etc).

2. Trabajos que se centran en el diseño de un almacén de datos, pero que no expresa una metodología para ello.

Para seleccionar los trabajos de investigación, en primera instancia utilizamos el criterio de inclusión para hacer análisis sobre el *título, resumen y palabras claves*, obteniendo de esta manera el mayor número de trabajos que aportan contribuciones significativas sobre los paradigmas para el diseño multidimensional de AD. En segunda instancia utilizamos el criterio de exclusión donde nos centramos principalmente en el resumen, introducción y conclusiones, analizando un poco más aquellos trabajos que lo requerían para asegurarnos realmente de que eran relevantes para el campo de estudio.

3.2 Selección de los estudios

Los estudios primarios proporcionan pruebas directas acerca de las preguntas de investigación. El proceso de selección consta de tres iteraciones llevadas a cabo por cuatro revisores. En la primera iteración, cada parte se examinó de manera independiente. Cada revisor aplicó los criterios de inclusión para cada trabajo, basado en el título, resumen y palabras claves. En la siguiente iteración los artículos considerados por los revisores fueron examinados nuevamente, ahora incluyendo la introducción y conclusión. En la tercera iteración se procedió a analizar completamente los trabajos que aún no convencían a los revisores.

3.3 definición del esquema de clasificación

Una vez seleccionadas las publicaciones relevantes se definieron, en base a los objetivos de estudios, cuatro tipos de clasificaciones (ver figura 2):

- ***Paradigma desarrollado****:* modelo en el cual están basados los artículos.
- ***Tipo de contribución****:* el aporte que realiza la investigación al área.
- ***Ámbito de aplicación****:* el área donde se desarrolla la investigación, o donde apuntan los autores para aplicar su investigación.
- ***Etapa de diseño****:* fase del desarrollo de un AD en la cual se centran los autores para desarrollar su investigación.

Cuando el esquema fue finalmente establecido, todas las publicaciones fueron revisadas nuevamente.

La clasificación según el paradigma desarrollado se realizó en tres categorías: *enfoque impulsado por la oferta, enfoque impulsado por la demanda, enfoque híbrido*.

El tipo de contribución fue clasificado en cuatro categorías: *Aproximación, metodología, herramienta, metodología-herramienta*. *Aproximación* para aquellas publicaciones que proponen nuevas ideas o metodologías distintas a las ya establecidas. *Metodología* incluye descripciones y procedimientos a seguir para realizar el diseño multidimensional de un AD. *Herramienta* se refiere a cualquier tipo de herramienta que ayude en el proceso de diseñar un AD. Por último *metodología-herramienta* para clasificar aquellas publicaciones que contribuyen con herramientas para aplicar la metodología presentada.

Ámbito de aplicación es clasificado en dos categorías: *Academia, industria*. *Academia* esta categorización es para aquellas publicaciones que dirigen su esfuerzo en realizar nuevas investigaciones y/o desarrollo de nueva ideas. *Industria* en este estudio esta clasificación corresponde a los trabajos que aplican su investigación en alguna organización (con o sin fines de lucro).

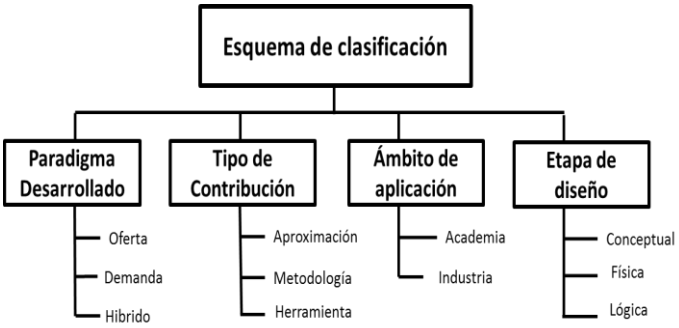

Figura 2. Esquema de clasificación

IV. Extracción de datos y mapeo sistemático

Tras definir el sistema de clasificación, el último paso del mapeo sistemático consiste en la extracción de datos y el proceso de mapeo de las distintas dimensiones. El resultado completo de esta actividad se muestra en la siguiente sección. El resultado sintetizado de nuestro estudio se puede observar de manera gráfica en el diagrama de burbuja de la figura 3.

La figura 3 ilustra básicamente dos diagramas de dispersión XY con burbujas en las intersecciones de categoría, que permite tener en cuenta varias categorías al mismo tiempo y da una visión general rápida de un campo de estudio, proporcionando un mapa visual. En esta visualización de los resultados, el tamaño de una burbuja es proporcional al número de artículos que están en el par de categorías que correspondan a la burbuja de las coordenadas.

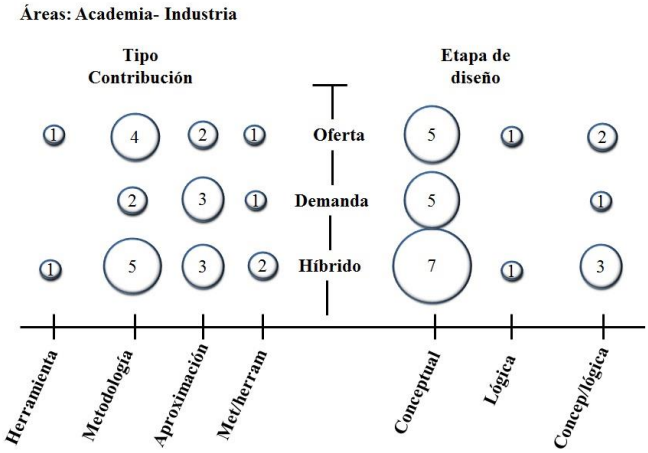

Figura 3. Diagrama de burbuja. Visualización mapeo sistemático

De igual forma, en la figura 4 se puede observar la distribución de trabajos por tipo de publicación y por enfoque de investigación. Del total de publicaciones incluidas en el mapeo (25), 11 son artículos, 9 pertenecen a revistas, 4 libros y 1 tesis.

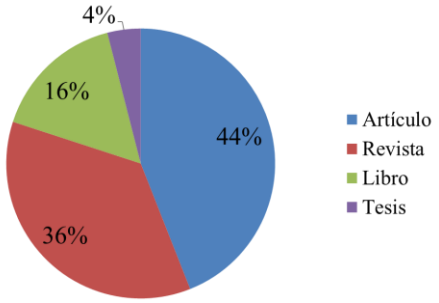

Figura 4. Distribución de los trabajos de investigación según tipo de publicación

4.1 Análisis comparativo y discusión

Luego de haber extraído y analizado la información relevante de cada estudio, presentaremos un resumen de algunos de los estudios que seleccionamos. La mayoría de los resúmenes se encuentran disponibles en el siguiente artículo: Estudio cronológico de paradigmas para el diseño de almacenes de datos del año 2012 [13].

Enfoque impulsado por la oferta

Mikael Jensen et al. [14] presentan una aproximación para la construcción automática del esquema de base de datos multidimensionales desde base de datos relacionales, permitiendo de manera fácil el diseño y análisis de la mayoría de las fuentes de datos. Esto lo logran a través de un conjunto de algoritmos prácticos y efectivos para el descubrimiento del esquema multidimensional de bases de datos relacionales.

Paul Hernandez [15] presentan una aproximación dirigida por modelos para obtener un modelo conceptual multidimensional desde los registros de datos de la Web de una manera comprensiva, integrada y automática. Esto se logra a través de la obtención del modelo conceptual de los registros web basados en un metamodelo unificado y derivando un modelo multidimensional de este modelo de registro web mediante la definición formal de un conjunto de reglas de transformación QVT (Query/View/Transformation).

Enfoque impulsado por la demanda

Bodo Hüsemann et al. [16] presentan una aproximación para obtener sistemáticamente un esquema conceptual del almacén de datos, que puede estar en una forma normal multidimensional (MNF) esto lo logran a través de la separación de las fases de diseño y analizando los requisitos para poder identificar los hechos, medidas y consultas habituales.

Ania Cravero et al. [17] presentan una aproximación en la cual se le da relevancia al análisis de la estrategia del negocio, la alineación entre los objetivos del almacén de datos y la estrategia de la empresa. Para lograr esto proveen un conjunto

de directrices que permite a los desarrolladores diseñar un AD alineado a la estrategia del negocio. La aproximación consiste en 4 fases: VMOST-basado en el análisis de la estrategia de negocio, analizando los elementos obtenidos usando BMM para alinear la estrategia de negocios con el AD, un modelo conceptual de un AD usando i* y un modelo multidimensional usando un perfil UML.

Enfoque hibrido

Romero y Abelló [18] presentan una aproximación para automatizar el diseño multidimensional de un AD, esto lo hacen a través de un método semiautomático que tiene como objetivo encontrar los conceptos multidimensionales del negocio a partir de fuentes de datos heterogéneas que no tienen nada en común, pero todas ellas son descritas por una ontología.

Glorio et al, [19] introducen información espacial en esta aproximación para cumplir con el desarrollo de un almacén de datos espacial (con información astronómica) usando MDA. Esto lo hacen a través de una extensión del nivel conceptual con elementos espaciales, también definiendo los principales artefactos MDA para modelar la información espacial en una vista multidimensional, además establecen formalmente un conjunto de reglas de transformación QVT para obtener automáticamente una representación lógica adaptada a la tecnología de una base de datos relacional. Por último aplicando las reglas de transformaciones QVT usando la herramienta MDA (basada en Eclipse) desarrollada para implementar el AD espacial.

Di Tria et al. [20] proponen una metodología híbrida secuencial para el diseño multidimensional de un AD, que tiene en cuenta tanto las ventajas del enfoque impulsado por la oferta a través del modelado de datos, como el Modelo de hechos dimensional (DFM) y las ventajas del enfoque impulsado por la demanda con la fuerte formalización de los requerimientos de los usuarios. La formalización es representada a través de UML, el cual posee un alto nivel de estandarización y una representación formal de los conceptos multidimensionales.

Lindsay Gómez et al. [21] proponen una metodología para el análisis de la información utilizando el paradigma hibrido intercalado, en donde primeramente analizan los requisitos de los usuarios para determinar las necesidades del usuario final y luego utilizan un análisis a través de las fuentes de datos. Se presentan pautas metodológicas para 4 etapas: análisis de la información, modelo conceptual, diseño lógico y trazabilidad. Estas etapas describen cómo se pueden obtener a partir de los sistemas operacionales heredados (E/R) una propuesta de modelado de almacenes de datos.

Thenmozhi y Vivekanandan [22] se enfocan en el diseño del esquema de un AD, para esto proponen una herramienta ontológica para automatizar el diseño del esquema multidimensional para el AD, las etapas del método son primero: la representación formal de los requerimientos, segundo derivando automáticamente elementos multidimensionales presentes en la fuente de datos ontológica, tercero alineando los requerimientos con la fuente de datos para filtrar los resultados, cuarto la generación de un esquema lógico, por ultimo aplicar la herramienta OBDWSD (Ontology based data warehouse schema deisgn).

A continuación damos respuesta a las preguntas de investigación formuladas en la sección 3 a través de los resultados obtenidos

- **PI1:** La oferta tiene 8, demanda tiene 6 y el hibrido 11. En cuanto a el paradigma que tiene más presencia, los números indican que la mayoría de las investigaciones actuales se han dirigido en el enfoque *híbrido* con 11 publicaciones de las 25(44%) seleccionadas. El segundo paradigma con mayor presencia es el enfoque impulsado por la *oferta* con un total de 8 trabajos (32%). Por otro lado el paradigma con menos presencia es el enfoque impulsado por la *demanda* con 8(24%) presencias. La tendencia en cuanto a la selección del paradigma por parte de la comunidad científica, es la preferencia por el enfoque híbrido para desarrollar sus trabajos. Un punto importante para mencionar es que el enfoque impulsado por la oferta sufrió un abandono desde el año 2004 sin embargo el año 2010 se volvió a retomar pero enfocándose en fuentes de datos de la web y lenguaje XML. En la figura 5 se muestran los resultados en medida porcentual.

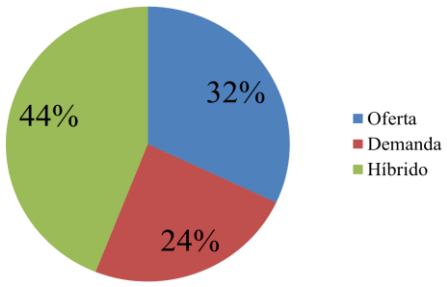

Figura 5. Distribución de las publicaciones según los paradigmas para el diseño multidimensional de AD

- **PI2:** En cuanto al ámbito en donde se aplican la mayoría de las contribuciones (*aproximaciones, metodologías y herramientas*), los números indican que la *industria* con 15 publicaciones, es el área donde la mayoría de los autores se enfoca para realizar y aplicar sus investigaciones. Por otro lado en el ámbito académico encontramos 10 publicaciones. En la figura 6 se muestra un gráfico con los resultados.

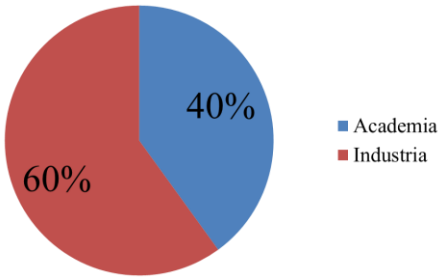

Figura 6. Distribución de los ámbitos en los que se enfocan las publicaciones

- **PI3:** Con respecto a la contribución que hacen los trabajos de investigación al área de diseño multidimensional de ADs, podemos observar en la figura 7 que la metodología resulta ser el aporte predominante en las investigaciones actuales con un total de 11 trabajos. El segundo lugar es para la aproximación con un total de 8 publicaciones, en tercer lugar encontramos una combinación entre metodología y herramienta con un total de 4 publicaciones. Este tipo de contribución combinado se debe a que los autores presentan en sus trabajos metodologías, que para poder aplicarlas lo hacen a través de una herramienta que ellos mismos crearon o modificaron de acuerdo a sus necesidades. También existen trabajos cuyo objetivo es presentar una herramienta y para este propósito presentan una metodología. En cuarto lugar se ubica la herramienta con 2 publicaciones.

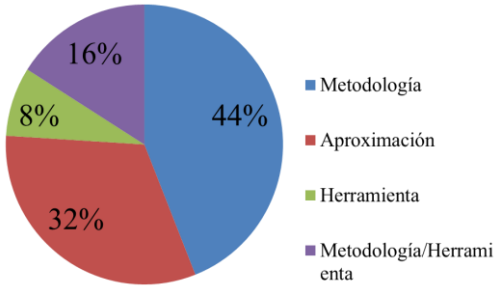

Figura 7. Distribución del tipo de contribución de las publicaciones al área

- **PI4**: En la figura 8 se puede apreciar que la etapa más investigada es la conceptual con un total de 17 publicaciones, esto es lo que se esperaba puesto que nos enfocamos en el diseño multidimensional, más que en las otras etapas. En segundo lugar encontramos publicaciones que están enfocadas en dos etapas: diseño conceptual y lógico, 6 son los trabajos que investigan esas dos áreas en conjunto. Por último encontramos 2 publicaciones orientadas a la etapa de diseño lógico.

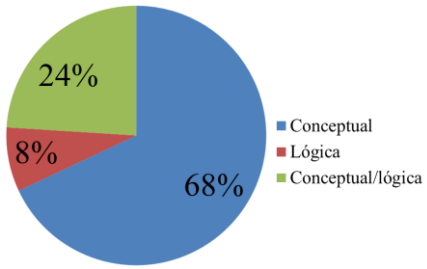

Figura 8. Distribución de las etapas de diseño en las que se enfocaron las publicaciones

4.2 Limitaciones del estudio

La principal limitación de este mapeo sistemático es haber analizado sólo 588 publicaciones del total de 2207 resultados obtenidos de las consultas realizadas en los diferentes buscadores. Aunque esta decisión sin duda perjudica la cobertura del estudio, creemos que esta desventaja se ve paliada por el uso de los buscadores más relevantes del campo (Google scholar, IEEE y Springer) y por la ordenación que hacen dichos buscadores de los resultados en función de su relevancia para la búsqueda. Cabe mencionar que puede que algunos documentos que existan no se hayan incluido, aunque la amplia revisión que se desarrolló y el conocimiento de este tema nos ha llevado a la conclusión de que, si existen, probablemente no son muchos.

La segunda limitación, debida a la propia filosofía del mapeo sistemático, es la calidad de los estudios incorporados. Esta limitación se podría haber evitado realizando una revisión sistemática de dichos trabajos. Sin embargo, el bajo número de publicaciones verdaderamente pertinentes (25), nos hace pensar que es todavía pronto para este tipo de evaluaciones de calidad. Otra posible limitación de este tipo de estudios es la posibilidad de errar en la clasificación por el uso ambiguo que hacen los autores como, en nuestro caso, aproximación o metodología. Relacionado con esto, hemos ratificado lo que ya advertían otros autores [23] acerca de los resúmenes de las publicaciones, que a menudo son engañosos y carecen de información relevante. En este estudio se ha limitado el impacto de estos riesgos mediante una aproximación conservadora al proceso de inclusión/exclusión de estudios, que ha implicado la lectura de cuantas partes de la publicación hayan sido necesarias hasta poder resolver la duda de si incluir o excluir el estudio.

V. Conclusiones

En este artículo hemos presentado un mapeo sistemático de estudios de los paradigmas para el diseño multidimensional de ADs, proporcionando un marco de trabajo actualizado, lo que nos permite formular nuevas actividades de investigación. El marco de revisión y el protocolo utilizados para la realización de esta revisión nos garantiza la completitud de los resultados. Como conclusión, la carencia más importante que hemos identificado es la falta de retroalimentación al momento de aplicar las metodologías, debido a que desarrollar, poner en marcha y obtener resultados en un AD demora mínimo 5 años aproximadamente aún no hay un artículo que trate sobre el proceso completo de la implementación de un AD ya sea en un ámbito académico o industrial. Otra carencia que identificamos es la falta de experimentos en esta línea de investigación, no hay publicaciones que traten sobre el diseño y desarrollo de un AD aplicando y comparando distintas metodologías existentes en el área. La falta de herramientas es otro punto que debemos destacar, en la mayoría de las publicaciones que contribuyen con herramientas, solo encontramos prototipos. No hay un instrumento finalizado y validado por los investigadores en el campo de investigación, esta carencia es importante considerando que la mayoría de las publicaciones se enfocan en la academia, mercado que determina la evolución del área. Además no hay una herramienta finalizada para realizar el diseño conceptual de manera automática, que es el objetivo que la mayoría de los autores quieren lograr.

Con el mapeo sistemático presentado en este trabajo se han conseguido identificar los principales trabajos de investigación sobre los paradigmas para el diseño multidimensional del AD publicados hasta el momento en los principales foros científicos. Este tipo de estudio permitirá facilitar la apertura del campo a nuevos investigadores.

5.1 Trabajos futuros

Se podría realizar una revisión sistemática debido al bajo número de publicaciones encontradas (25), esta metodología es parecida al mapeo sistemático, sin embargo la fase de revisión de trabajos es mucho más rigurosa, permite establecer el estado de evidencia a través de la exhaustiva extracción de datos cuantitativos y estudios e meta-análisis, y por tanto responder a preguntas de investigación mucho más específicas [24]

VI. Agradecimientos

Este trabajo fue "Financiado (parcialmente) por Dirección de Investigación, Universidad de La Frontera".

REFERENCIAS

[1] J. N. Mazón, Trujillo, J., Serrano, M., Piattini, M., "Designing Data Warehouses: From Business Requirement Analysis to Multidimensional Modeling.," REBNITA'05, pp. 44-53, 2005.

[2] P. Giorgini, S. Rizzi, and M. Garzetti, "GRAnD: A goal-oriented approach to requirement analysis in data warehouses.," Decision Support Systems., vol. 45, pp. 4 - 21, 2008.

[3] R. Kimball and M. Ross, "The Data Warehouse Toolkit, second edition, John Wiley & Sons.," 2002.

[4] J. N. Mazón and J. C. Trujillo, "Desarrollo de modelos multidimensionales de almacenes de datos basado en MDA: del análisis de requisitos al modelo lógico," 2007.

[5] R. Winter and B. Strauch, "Information Requirements Engineering for Data Warehouse Systems," 2004.

[6] B. Kitcheham, T. Dyba, and M. Jorgensen, "Evidence-based software engineering, in Proceeding of the 26th Int. Conf. on Software Engineerig(ICSE, 2006), IEEE Computer Society.," pp. 373-378, 2006.

[7] W. Inmon, "Building the Data Warehouse.," 2005.

[8] M. Golfarelli and S. Rizzi, "A methodological framework for data warehouse design.," Proceedings DOLAP'98., pp. 3-9., 1998.

[9] E. Codd, S. Codd, and C. Salley, " Providing OLAP to user-analysts: An IT mandate., E. F. Codd and Associates.," vol. 32, 1993.

[10] R. Kimball, L. Reeves, W. Thornthwaite, and M. Ross, " The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses.," John Wiley & Sons, Inc., 1998.

[11] R. Winter and B. Strauch, "A method for demand driven information requirements analysis in data warehousing projects., Proceedings of the 36th Annual Hawaii International Conference on.," System Sciences, pp. 9-14, 2003.

[12] O. Romero and A. Abello, "A survey of Multidimensional Modeling Methodologies.," International Journal of Data Warehousing & Mining., vol. 5, nº. 2, pp. 1-23, 2009.

[13] A. Cravero and S. Sepúlveda, "A chronological study of paradigms for data warehouse design," Ingeniería e Investigación, vol. 32, nº. 2, pp. 58-62, 2012.

[14] J. Jensen, T. Holmgren, and T. Pedersen, "Discovering multidimensional structure in relational data," Data Warehousing and Knowledge Discovery, pp. 138--148, 2004.

[15] P. Hernandez, I. Garrigos, and J.-N. Mazon, "modeling web logs to enhance the analysis of Web usage data," Database and Expert Systems Applications (DEXA), 2010., pp. 297--301, 2010.

[16] B. Hüsemann, J. Lechtenbörger, and G. Vossen, "Conceptual Data Warehouse Modeling. In M. A. Jeusfeld, H. Shu, M. Staudt, G. Vossen (Eds.)," Proceedings of 2nd International Workshop on Design and Management of Data Warehouses., pp. 6, 2000.

[17] A. Cravero, J.-N. Mazón, and J.Trujillo, "A business-oriented approach to data warehouse development," Revista Ingeniería e Investigación, vol. 33, nº. 1, pp. 59-65, 2013.

[18] O. Romero and A. Abelló, "Multidimensional Design Methods for Data Warehousing," Integrations of Data Warehousing, Data Mining and Database Technologies: Innovative Approaches, pp. 78, 2011.

[19] O. Glorio, J.-N. Mazón, I. Garrigos, and J. Trujillo, "Using web-based personalization on spatial data warehouses. Proceedings of the 2010 EDBT/ICDT Workshops," pp. 8, 2010.

[20] F. Di-Tria, E. Lefons, and F. Tangorra, "Hybrid methodology for data warehouse conceptual design by UML schemas," Information and Software Technology, vol. 54, nº. 4, pp. 360--379, 2012.

[21] L. Gómez, R. Moreno, and R. Pérez, "Computer - Asssted generation of data warehouse model: analysys of information," DYNA, vol. 80, nº. 177, pp. 49--58, 2013.

[22] M. Thenmozhi and K. Vivekanandan, "An Ontology based Hybrid Approach to Derive Multidimensional Schema for Data Warehouse," International Journal of Computer Applications, vol. 54, nº. 8, pp. 36--42, 2012.

[23] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," EASE'08 Proceedings of the 12th international conference on Evaluation and Assessment in Software Engineering. British Computer Society Swinton., pp. 68-77, 2008.

[24] Meliá, Santiago, Cristina Cachero, and Yulkeidi Martínez. "Evidencia empírica sobre mejoras en productividad y calidad en enfoques MDD: un mapeo sistemático." REICIS Revista Española de Innovación, Calidad e Ingeniería del Software 7.2 (2011): 6-27.

[25] Kitchenham, Barbara A., and Stuart Charters. "Guidelines for performing systematic literature reviews in software engineering." (2007).

[26] Bailey, J., Budgen, D., Turner, M., Kitchenham, B., Brereton, P., & Linkman, S. (2007, September). Evidence relating to Object-Oriented software design: A survey. In Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on (pp. 482-484). IEEE.