Data Warehousing en soluciones de Inteligencia de Negocios

Lucina García Hernández[1] (Facultad de Matemática y Computación, Universidad de La Habana, Cuba / Lis Velázquez Vidal[2] (División de Sistemas de Gestión Empresarial DATYS, Cuba) / Mijail Veliz Monteagudo[3] (División de Turismo DATYS, Cuba)

Resumen: la Inteligencia de Negocios constituye una estrategia de trabajo para el desarrollo de soluciones computacionales que transforman los datos primarios en conocimiento en función de la misión y la visión de una institución. Un data warehouse es una herramienta de carácter universal en el contexto organizacional que facilita la comprensión de los datos al transformarlos en información útil de modo que sea posible crear conocimiento.

La población de un data warehouse consiste en la aplicación de un conjunto de técnicas que provee un soporte comprensible para la copia y la transformación, esencialmente automática, de los datos desde las locaciones de origen hasta la ubicación de destino de manera controlada, consistente, repetible y viable. En la actualidad existen diversas herramientas orientadas al data warehousing sobre plataformas particulares como soporte de los datos; no obstante, generalmente no están comprometidas con el proceso de población en todo su alcance.

El presente artículo tiene como objetivo contribuir al diseño y la instrumentación del data warehouse con un enfoque arquitectónico de tres capas, que proporcione alternativas para su propio enriquecimiento así como para la evolución del entorno en que se desarrolla. Se incluye un procedimiento para el diseño del flujo ETCL así como una formalización matemático computacional para el modelo de datos multidimensionales. Se expone un modelo de solución genérico para representar el proceso de creación de estructuras multidimensionales a partir de un almacén único de datos, que propicia la creación de herramientas extensibles e independientes de los gestores de bases de datos.

1. INTRODUCCIÓN

Para cualquier empresa el desempeño del negocio es la realización de una serie de eventos cuya naturaleza y frecuencia varían según las funciones particulares de la organización. De cualquier manera los eventos siempre están presentes y constituyen la base y el componente fundamental de la marcha del negocio. Mantener un registro de los eventos que garantice la información detallada, generada como parte de la interacción entre los eventos y los procesos de la empresa, constituye un ingrediente fundamental en la toma de decisiones.

La información —cada día más abundante y diversa, procedente de múltiples fuentes en diferentes formatos y que es imprescindible recoger, ordenar y manipular para obtener un valor añadido— forma parte de la habilidad competitiva de las organizaciones. El uso de la información como un arma estratégica, sustentado por modelos matemático computacionales como soporte de aplicaciones analíticas que maximicen el rendimiento de las organizaciones en sus actividades para generar eficiencia operativa y asegurar la evolución, constituye una necesidad en el mundo actual.

Uno de los enfoques más difundidos en la actualidad consiste en la explotación inteligente de los recursos computacionales con vistas a favorecer el accionar y el devenir de la organización.

[1] lucina@matcom.uh.cu
[2] lis.velazquez@datys.cu
[3] mijail.veliz@datys.cu

Corresponde, entonces, buscar respuestas a la necesidad de garantizar el funcionamiento y el progreso de la empresa mediante la combinación apropiada de lo existente con los nuevos paradigmas y tecnologías en aras del enriquecimiento del escenario.

La clave del uso del paradigma de la Inteligencia de Negocios en cualquier organización, sea o no lucrativa, radica en el tratamiento de la información. En la actualidad, la mayoría de las grandes compañías productoras de software —Microsoft, Oracle, SAS Institute, Hyperion Solutions, entre otras— ofrecen herramientas para la construcción de soluciones de inteligencia de negocios que incluyen utilidades para data warehousingy el procesamiento analítico sobre plataformas particulares como soporte de los datos, aunque no están comprometidas con el proceso de población en todo su alcance [7].

El presente artículo pretende contribuir al diseño y la instrumentación del data warehouse con un enfoque arquitectónico de tres capas, que proporcione alternativas para su propio enriquecimiento a partir de la aportación del modelo dimensional para representar el escenario organizacional, así como para la evolución del entorno en que se desarrolla dado el carácter iterativo de las soluciones de inteligencia de negocios. Se incluyen los fundamentos para la confección de unambiente computacional extensible y genérico orientado a la conformación del proceso de población del data warehouse.

2. LOS DATOS ESTRUCTURADOS EN LAS SOLUCIONES DE INTELIGENCIA DE NEGOCIOS

Teniendo en cuenta que coexisten numerosas definiciones más o menos coincidentes, los autores hemos considerado provechoso insistir en que "La Inteligencia de Negocios se puede resumir como una estrategia de trabajo que persigue optimizar la toma de decisiones apoyándose en un conjunto de instrumentos y técnicas enfocadas a la administración y la extracción de conocimiento mediante el análisis de los datos existentes en una organización con el fin de contribuir certera y pertinentemente al éxito empresarial" [7].

Se trata de modelar, diseñar e implementar una plataforma matemático-computacional con un enfoque integrador que contribuya a mantener actualizados los datos, procesar y analizar la información diaria e histórica, realizar diagnósticos, pronósticos y revelaciones para tomar decisiones y proyectar acciones tácticas y estratégicas en función de los objetivos de la organización.

Resulta más procedente hablar de sistemas o soluciones de Inteligencia de Negocios puesto que no existe un modelo único para su desarrollo. Una solución de Inteligencia de Negocios no solo consiste en interpretar los datos y los procesos de la empresa en términos de la toma de decisiones, sino también la concepción de una arquitectura matemático-computacional que permita combinar y desarrollar creativamente un conjunto de herramientas para instrumentar un ambiente de trabajo orientado a los especialistas y los ejecutivos de la institución.

A partir de la diversidad y la pluralidad del análisis y la contribución específica de las herramientas se requiere construir la solución sobre la base de aproximaciones, mediante la aplicación de un enfoque evolutivo que la conduce a un estatus superior en cuanto a la toma de decisiones a medida que se trabaja y se lanza nuevas estrategias. Por otro lado, asegurar la calidad de los datos en el sentido de su existencia,validez, exactitud y veracidad también constituye un aspecto trascendental en las soluciones de Inteligencia de Negocios.

Sin dudas, la arquitectura de las soluciones de inteligencia de negocios ha progresado. En estas soluciones no basta con tener un manejador de datos robusto, ya que subsiste la problemática de la extracción de datos almacenados en otros sistemas. Una visión contemporánea contempla una infraestructura de middleware que habilite y gestione la comunicación para la integración entre componentes aplicativos diversos [13].

En realidad no es imprescindible la presencia de un data warehouse (DW) en una solución de Inteligencia de Negocios, ya que los escenarios pueden ser heterogéneos. Sin embargo, es preciso tomar en consideración la contribución de los datos factuales para evaluar el progreso de una organización así como el propósito que persigue el proceso de población del data warehouse, en cuanto a la integración de los datos, con vistas a la exploración multidimensional acertada.

En estudios anteriores [1, 17] se afirmó y se comprobó que un data warehouse está orientado a optimizar la toma de decisiones, razón por

la cual muchas de las soluciones de Inteligencia de Negocios incluyen el proceso de data warehousing, puesto que contempla la administración de los datos estructurados y la información desde su origen y brinda un entorno favorable para la extracción de conocimiento.

Existe la opinión que no siempre un data warehouse es apropiado para la aplicación de algoritmos de la minería de datos. Aún así, durante las diferentes etapas del proceso de creación y mantenimiento deldata warehouse se realizan operaciones que pueden ser aprovechadas o combinadas con las requeridas para la obtención y la generación del conocimiento [6].

La definición de data warehouse más aceptada por la comunidad científica fue expuesta en la década de 1990 por William H. Inmon quien planteó que: "Un data warehouse es una colección de datos integrada, orientada a sujetos, variante en el tiempo y no volátil, utilizada como apoyo para los procesos de toma de decisión." [8].

Para todo data warehouse es una necesidad ser suficientemente flexible ante cambios en los datos así como en las exigencias de análisis de la información. En este sentido las implementaciones particulares responden a arquitecturas de los datos que se diseñan de acuerdo a la complejidad conceptual del fenómeno y cada autor propone una nomenclatura para denominar las capas del data warehouse. Asimismo, la existencia de las capas es también un tema polémico.

Dos de los exponentes fundamentales del data warehousing asumen posiciones distintas. Para Inmon [9], el data warehouse está formado por los datos reconciliados mientras que los datos derivados o vista informacional están fuera desu alcance. Con este criterio se está mutilando al almacén de datos de su carácter informacional destinándolo solo a una función reguladora.

En el otro extremo conceptual se encuentra Ralph Kimball [10], quien expone que el data warehouse corresponde a la unión de varias representaciones multidimensionales de los datos divididos por temáticas. Esta opinión no tiene en cuenta que siempre que se colectan datos de diversas fuentes es necesaria la transformación y la depuración para lograr la integración.

Analizando estos puntos de vista extremos se refleja la importancia que debe tener cada etapa dentro del proceso. Una posición intermedia se refiere a la arquitecturaconceptual de los datos de tres capas propuesta por Devlin [5],cuya nomenclatura se expone a continuación.

- Los sistemas operacionales como origen de los datos de tiempo real conforman la primera capa lógica, que se refiere a los datos primarios en sus soportes respectivos.
- El data warehouse empresarial (DWE) que almacena fisicamente los datos reconciliados de las distintas fuentes operacionales, creando una vista única de los datos de toda la organización como respuesta a la integración, constituye la segunda capa lógica.
- El warehouse informacional (WI) que conserva los datos bajo un enfoque dimensional que permite múltiples derivaciones, resumiéndolos y optimizándolos para las consultas desde perspectivas analíticas diferentes, asienta la tercera capa.

Podría pensarse que este enfoque es impracticable por el tiempo y los recursos que demanda. Es cierto que, en ocasiones, existen sistemas operacionales que reflejan el universo completo del fenómeno a modelar y almacenan datos históricos. Sin embargo, la práctica ha demostrado ser más rica de lo que sea posible imaginar y lo que parece suficiente hoy, mañana pudiera no serlo. Teniendo en cuenta el surgimiento de fuentes de datos no consideradas en un primer diseño —tal vez por no existir—, se sugiere como paradigma y se adopta en el marco de la presente investigación la arquitectura conceptual de tres capas. Análisis más detallados al respecto se realizan en [17, 18].

Por otro lado, asegurar la calidad de los datos en el sentido de que sean precisos, consistentes, completos y sin ambigüedades constituye un aspecto trascendental en las soluciones de Inteligencia de Negocios. Indudablemente, la detección y la solución automática de errores es un proceso que reviste alta complejidad y, por lo regular, entraña el establecimiento de compromisos entre los desarrolladores y los usuarios finales.

El proceso de extracción, transformación y carga (ETL) de los datos originales en un data

warehouse constituye también un contexto apropiado para profundizar en esta arista del problema dados sus objetivos de integración y expresión cualitativa a través del tiempo y según la planificación prevista. Consecuentemente, la mejor opción para el logro de estos propósitos es precisamente la inclusión de almacenes de datos en su concepción más amplia en el marco de las soluciones de Inteligencia de Negocios.

La creación y la explotación de un data warehouse puede contar con varias etapas de población que, aun cuando tienen un estrecho vínculo, son modularmente diferenciables en dependencia de la arquitectura seleccionada. En la arquitectura de tres capas existen dos procesos de población, a saber, la población del data warehouse empresarial desde los sistemas operacionales y la población del warehouse informacional desde el data warehouse empresarial.

Cabe destacar que en gran parte de la bibliografía el proceso ETL utilizado para data warehousing incluye tanto la población, la integración y la limpieza de los datos como la creación de los datos informacionales. Claro está que, en la arquitectura de tres capas los procesos de población del DWE y del WI se independizan, pues cumplen objetivos distintos y bien delimitados.

En el presente tema de investigación se ha profundizado en el proceso ETL con el propósito de establecer un procedimiento que ha servido de base a la concepción y la elaboración de una herramienta cuyas características esenciales responden a la diversidad funcional y estructural del proceso de población de un data warehouse empresarial así como a brindar flexibilidad para su implantación en diferentes ambientes computacionales.

Asimismo, se ha penetrado en el modelo dimensional en la arista teórico-conceptual al incursionar en una formalización matemática, en el ángulo metodológico al profundizar en el diseño multidimensional y en el ámbito aplicativo al servir de fundamento para la concepción de una herramienta que permite implementar y enriquecer el proceso de población del warehouse informacional en diversas plataformas.

Los procesos de población propios de un data warehouse tienen una relación de interdependencia que no se manifiesta solo desde el punto de vista de diseño, sino también de explotación. De ahí que, para obtener un resultado exitoso en función de garantizar la calidad de los datos y, por consecuencia, la utilidad de la información puesta a disposición de los usuarios finales, sea imprescindible la interacción entre los procesos de población.

Por ende, ambas herramientas no solo son extensibles y genéricas sino que se complementan para brindar un entorno que favorece la instrumentación del proceso de población de un data warehouse así como de los data marts respectivos que comprendan las aproximaciones sucesivas de la solución de Inteligencia de Negocios para una organización según las prioridades informacionales.

3. EL PROCESO DE POBLACIÓN DEL DATA WAREHOUSE EMPRESARIAL

El diseño físico correcto y el control del proceso de población determinan no solo la estructura adecuada de los datos puestos a disposición para el análisis y la toma de decisiones de los usuarios finales sino también su autenticidad y oportunidad. Entre los objetivos del data warehouse empresarial se encuentra dar respuesta al carácter integrador y a la variación en el tiempo de los datos de un data warehouse.

El proceso de población del data warehouse empresarial es el encargado de convertir y unificar los datos operacionales en un único repositorio de datos con un enfoque relacional. Por tanto, la copia de los datos en el DWE constituye un tipo especial de réplica, en el cual se modifican los datos operacionales capturados para obtener un escenario más completo y consistente de la organización desde la perspectiva de los datos.

Entre las tareas del DWE se encuentra la conciliación de los datos de diversos orígenes. En muchos casos estos datos difieren en el diseño de sus estructuras, carecen de llaves primarias o poseen llaves que dificultan su manipulación. Para resolver estos problemas se propone el empleo de las llaves sustitutas (surrogatekeys).

Las llaves sustitutas son enteros asignados de forma secuencial por el diseñador del DWE, según sea necesario para poblar un conjunto de entidades. Estas llaves no tienen significado empresarial e identifican inequívocamente cada

elemento del conjunto. Sin embargo, por simple que pudiera parecer la solución, la administración de estas llaves es bien compleja pues es preciso preservar la integridad referencial para garantizar el funcionamiento correcto del data warehouse.

El empleo de llaves sustitutas asegura la optimización de las consultas, la depuración de los datos operacionales y apoya el mantenimiento de la historia. Desempeñan un papel fundamental en el procesamiento de los cambios en el modelo informacional, puesto que permiten diferenciar las entidades originales de las que se generan con el objetivo de conservar el registro de las modificaciones en los datos y sus estados a través del tiempo con vistas a suministrar la información requerida parala realización de análisis históricos y la predicción del comportamiento de la organización.

Es preciso también garantizar la fidelidad de los datos que tributan al análisis aunque no sea imprescindible mantener el registro del momento justo en que se realizan los cambios. Vale acotar que los mecanismos de solución que se incluyen en este epígrafe, aunque se enfocan desde el punto de vista dimensional, son válidos para el data warehouse empresarial porconstituir la base delwarehouse informacional.

La determinación de la frecuencia de variación de las dimensiones es relativa pero se han propuesto métodos combinados para almacenar los cambios, subdividiendo las dimensiones de forma que los atributos que se actualizan con mayor asiduidad conformen un conjunto de entidades independiente para optimizar el almacenamiento de las modificaciones y sea posible aplicar otras prácticas a las dimensiones de cambio lento,también conocido por sus siglas SCD (del inglés, Slowly Changing Dimension) [12].

El tratamiento básico propuesto para el enfoque dimensional como respuesta a los cambios en los datos fuentes se resume en sobrescribir el valor (Tipo 1), adicionar un registro (Tipo 2) y/o adicionar un campo (Tipo 3). El Tipo 1 es la implementación típica en los sistemas operacionales, pero no permite mantener el registro de los cambios en el atributo en cuestión.El Tipo 2 mantiene el vínculo con el elemento modificado en el ambiente operacional, así como con todos los estados anteriores, mediante el uso de llaves sustitutas. El Tipo 3 se aplica a menudo cuando, a pesar de haber tenido lugar un cambio, es aún lógicamente posible actuar como si el mismo no hubiera acontecido o referenciar los registros indistintamente por sendas versiones del atributo en cuestión.

Por otra parte, las marcas de tiempo (timestamps) cumplen una función fundamental en la persistencia de la historia en el data warehouse al soportar la representación de los datos periódicos y sus estados ya que consignan el momento en que un registro se inserta, se borra o se modifica. En una base de datos relacional las marcas de tiempo se adicionan a la llave primaria original de la tabla con el objetivo de identificar cada tupla inequívocamente y/o constituir la expresión física de los metadatos de temporalidad que pueden aplicarse a los conjuntos de entidades, las tuplas o los atributos.

El mecanismo más ampliamente difundido para reflejar los cambios de estado de los datos se basa en el uso de marcas dobles de tiempo de modo que los atributos identifican el inicio y el fin del período de prevalencia de cada registro. Este mecanismo garantiza la optimización de las consultas aunque adiciona una sobrecarga en la actualización, que es despreciable dada la orientación informacional del data warehouse.

En realidad, las marcas de tiempo y las SCD no son mecanismos mutuamente excluyentes, sino que pueden combinarse con facilidad aprovechando las llaves sustitutas para garantizar la integridad referencial, lo que resulta una alternativa conveniente para asegurar la calidad de los datos durante el proceso de población del DWE así como para la población y la explotación del WI.

Aun cuando el proceso de población de un data warehouse se divide en tres etapas esenciales: extracción, transformación y carga, algunos autores —entre los que se destaca Ralph Kimball [11]— redistribuyen el flujo de datos mediante una propuesta funcional aplicada específicamente al enfoque dimensional. Desde este punto de vista el proceso se dividiría en Extracción, Depuración, Integración y Distribución (una traducción lo más cercana posible de Extraction, Cleaning, Conforming and Delivery).

Al incluir una etapa independiente de depuración, esta nueva propuesta realza la impor-

tancia de garantizar la calidad de los datos almacenados en el data warehouse pero reduce la transformación a depuración e integración únicamente. De esta forma se está obviando uno de los objetivos fundamentales del data warehouse dado que durante el flujo de datos en el proceso de población se efectúan transformaciones con fines meramente informacionales que no intentan integrar los datos obtenidos de las distintas fuentes sino enriquecer la respuesta que se pone a disposición de los usuarios finales.

El esquema de la figura 1 constituye una propuesta de flujo de datos que incluye la Depuración (Cleaning) como una etapa independiente, aunque mantiene el término Transformación para describir de manera más abarcadora la etapa de transición de los datos desde el origen hasta la estructura adecuada en el destino final. En este trabajo preferimos abordar el proceso ETCL con respecto a cómo enfrentar su instrumentación.

En la extracción de los datos se debe tener en cuenta la variedad de formatos existentes en los sistemas operacionales, los que pueden no solo encontrarse almacenados en diferentes bases de datos relacionales, sino también en otros soportes menos estructurados como hojas de cálculo, ficheros XML o texto.

En la etapa de transformación los datos obtenidos de los sistemas operacionales deben modificarse con el fin de homogeneizar sus formatos para integrarlos consistentemente de acuerdo con la estructura y el diseño del DWE. Generalmente, estas transformaciones distan de la replicación clásica, por lo que se precisa de la combinación de operaciones específicas aplicadas a las fuentes durante el proceso de migración en dependencia del resultado deseado.

Entre las operaciones más frecuentes que se aplican para transformar el conjunto de datos fuente en el conjunto de datos destino se destacan laselección, la concatenación/separación, lanormalización/denormalización, la transposición, la agregación, la conversión y el enriquecimiento. En sentido general, las transformaciones son consideradas componentes de una etapa independiente en el flujo ETL, sin embargo, por lo regular están distribuidas a lo largo de todo el proceso de población.

La calidad de los datos constituye un aspecto importante para una gran parte de los usuarios de la información, no solo para grupos de usuarios específicos como los administradores y desarrolladores del data warehouse, sino también para los usuarios finales, ya que sin datos correctos es imposible pensar en tomar decisiones certeras.Los datos que se pondrán a disposición de los usuarios finales en el data warehouse deben ser correctos, lo que puede interpretarse en el sentido de que sean válidos, satisfagan las reglas del negocio y estén bien definidos[14].

La depuración de los datos constituye un tipo de transformación no solo por el hecho de modificar los datos sino porque para su implementación se utilizan los tipos de transformaciones expuestos con antelación. El principio básico de la etapa de depuración de errores es detectar los problemas de calidad en los datos, corregirlos automáticamente si es factible y en caso negativo registrar su ocurrencia, lo que debe llevarse a cabo evitando rechazar registros o detener el flujo ETCL.

Durante la carga los datos se almacenan con la estructura adecuada para servir de fuente al warehouseinformacional y se asegura el mantenimiento de la historia de los cambios ocurridos en los datos.

Las etapas de extracción, transformación, depuración y carga no son independientes pues

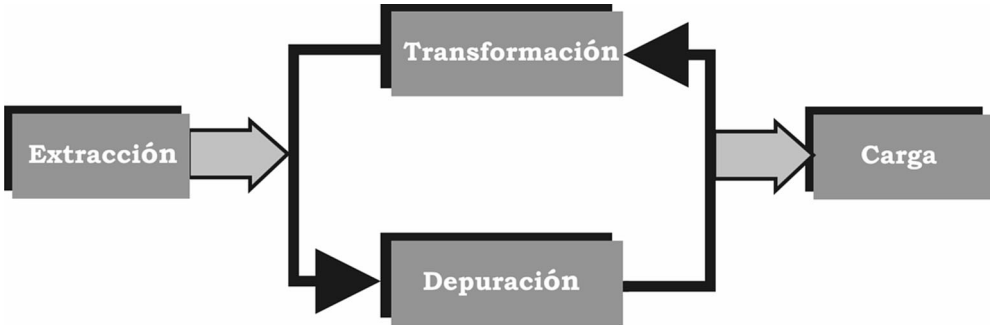

Fig. 1. *Etapas del proceso ETCL.*

la combinación de ellas guía el flujo de un conjunto de datos específico desde la fuente hasta el DWE. Con el objetivo de analizar el proceso ETCL y orientar su instrumentación definiremos sus componentes esenciales.

- Acción: consiste en una operación atómica con un significado lógico que puede representar la extracción de un conjunto de un origen dado, una transformación específica, un chequeo de calidad o una modalidad de carga.
- Tarea: agrupa un conjunto de acciones dependientes entre sí, representativo de una transformación lógica en los datos, que puede abarcar desde la extracción hasta la carga en el conjunto final del DWE.
- Subproceso: reúne un conjunto de tareas dependientes entre sí.Un subproceso puede supeditarse a la consumación de uno o varios subprocesos, puede ejecutarse en un punto predeterminado en el tiempo o puede ser desencadenado por un evento empresarial.
- Planificación: comprende el diseño y la realización de un plan consecuente en función de los subprocesos, su dependencia temporal o subordinación eventual.
- Ejecución: se encarga de cumplimentar las acciones, las tareas y los subprocesos a partir de la planificación diseñada.
- Notificación: consiste en la emisión de confirmaciones de ejecución o la divulgación de avisos para tomar las medidas que atenúen o eliminen los problemas detectados, manteniendo un registro del estado de las alertas.

Para diseñar el proceso ETCL es necesario determinar las operaciones lógicas con un comportamiento atómico, agruparlas en base a su interdependencia y programar la ejecución de las mismas en función del proceso en su totalidad.

Así tendremos que la etapa de extracción puede diseñarse como una acción lógica independiente que, parametrizada en función del origen y el formato de los datos, constituya una estructura atómica del proceso ETCL independientemente del mecanismo de captura de cambios utilizado.

De igual forma, cada uno de los tipos de transformaciones analizados puede ser tratado como una acción, cualquiera sea su complejidad.Cada uno de los chequeos de calidad constituye también una operación atómica que, en dependencia de los errores detectados, puede generar como resultado un conjunto de datos y, en función de su posible solución, modificar el conjunto resultante o dejarlo intacto.

Finalmente, es posible considerar los tipos de carga como acciones. Su funcionamiento varía en dependencia de la captura de cambios utilizada y del tipo de respuesta a los cambios para las SCD que se empleen en el conjunto de datos.

Habiendo definido cada operación atómica se podrá planificar las tareas a instrumentar. Cada una de las acciones seleccionadas, agrupadas y ordenadas adecuadamente en función de los datos que se desea poner a disposición de los usuarios finales conforman las tareas del proceso ETCL que existen con el objetivo marcado de consolidar acciones lógicas dependientes. Una tarea encierra una semántica pues reúne las acciones que guían un conjunto de datos o varios dependientes entre sí a través del flujo del proceso ETCL.

Las tareas también pueden ser agrupadas en subprocesos con una funcionalidad un tanto diferente. Los conglomerados de tareas dependientes entre sí existen para garantizar la ejecución apropiada de las mismas en el flujo del proceso ETCL. Los subprocesos forman parte de los componentes básicos del proceso ETCL ya que garantizan la integridad del proceso de población en función del tiempo.

Las implementaciones que excluyen la categoría de los subprocesos permiten, entonces, planificar las tareas. Esto ocasiona que a corto o largo plazo terminen sucumbiendo ante la necesidad de especificar la concurrencia de tareas predecesoras en función del tiempo, lo cual provoca redundancia entre las instancias de una misma tarea proyectadas en el tiempo.

Es importante tener en cuenta el orden y, en consecuencia, el paralelismo de las acciones, las tareas y los subprocesos para asegurar el funcionamiento correcto y la optimización del proceso ETCL durante la ejecución, cuidando no incurrir en los problemas clásicos —como el deadlock y los ocasionados por la concurrencia— que son complejos de detectar y depurar.

Por otra parte, la planificación de tareas o subprocesos debe estar en correspondencia con

el funcionamiento de la empresa de modo que no se afecte la actividad de los sistemas operacionales. Adicionalmente, es necesario contar con un subsistema de notificación que favorezca la inmediatez de la respuesta manual a los problemas que puedan detectarse durante la población del DWE.

De este modo se resume el procedimiento que ha servido de base a la concepción y la elaboración de la herramientaque facilitala instrumentación del proceso de población del data warehouse empresarial en correspondencia con la arquitectura de datos de tres capas.

4. RIQUEZA DEL ENFOQUE DIMENSIONAL

El warehouse informacional constituye la tercera capa de la arquitectura física del DW. En ella se almacenan los datos con vistas a su análisisulterior desde diferentes perspectivas.

La modelación es una manera gráfica y efectiva de representar los procesos y las reglas que caracterizan un fenómeno o negocio. La modelación empresarial abarca la representación sintética del escenario de la organización en cuanto al flujo de los procesos empresariales y la circulación de los datos para expresar el comportamiento dela institución y podría ser utilizada por sus directivos para desarrollar estrategias y trazar las líneas de acción.

La modelación de los datos persigue representar el universo de datos con que trabaja la organización y las relaciones entre ellos. A grandes rasgos se puede plantear una secuencia, no necesariamente obligatoria, para obtener un modelo de los datos de una empresa desde el punto de vista funcional, avalado computacionalmente por el enfoque relacional con un alto grado de normalización para minimizar las anomalías de actualización.

Ahora bien, si se desea caracterizar la realidad empresarial en términos analíticos resulta más apropiado emplear el enfoque dimensional. Al trabajar con la multidimensionalidad se persigue examinar los datos desde perspectivas diferentes con el fin de lograr una visión global de la problemática que permita fundamentar las decisiones estratégicas en diferentes circunstancias, objetivas y subjetivas, con una incidencia significativa de la temporalidad.

Una estructura multidimensional representa un paradigma de bases de datos que intenta reflejar de manera física varias dimensiones dado que el enfoque relacional solo aporta estructuras de dos dimensiones. De ahí que se introduzca el concepto de cubo de información, cuyas celdas contienen resúmenes de los datos, según múltiples aristas.

Los cubos son los objetos principales del procesamiento analítico en línea. Se conoce como cubo a los datos que se organizan y se resumen en una estructura multidimensional representada por un juego de medidas —que consiste en el conjunto de valores con que se mide el desempeño de la actividad que se esté analizando— y dimensiones —características o propiedades que brindan disímiles perspectivas de análisis sobre un hecho dado— que responden al fenómeno o proceso en cuestión. [16]

El enfoque multidimensional está dirigido a mejorar la eficiencia de las consultas. Por supuesto, la eficacia de los análisis multidimensionales depende de la manera en que los datos se representen y se almacenen. En cualquier caso, resulta esencial la presencia de los expertos en la actividad que se está modelando así como su estrecha interrelación con el diseñador para precisar las reglas del negocio y los requerimientos informacionales.

Los modelos matemáticos de datos constituyen en esencia el sostén orgánico de los sistemas de gestión de bases de datos, aun cuando los primeros SGBD convencionales aparecieron en la práctica antes que laformalización matemática de los enfoques utilizados para la representación computacional de los datos.

Los modelos matemáticos de datos se utilizan para el diseño lógico de los datos y se basan, sustancialmente, en conceptos que son abstracciones de las estructuras de las bases de datos y no abstracciones del universo de discurso. Los modelos matemáticos de datos están formados por tres componentes, a saber,

- Las estructuras de los datos para la representación de los datos y sus interrelaciones.
- Las reglas de integridad para delimitar los valores que pueden estar presentes en la base de datos.
- Las operaciones para la manipulación de los datos.

Al intentar procesar los datos para la toma de decisiones desde la perspectiva matemático-computacional se ha utilizado varios conceptos y técnicas. Entre ellos se pueden mencionar las ecuaciones o funciones matemáticas que describen el comportamiento de los datos, así como las reglas estadísticas que, además, facilitan la obtención de diagnósticos y pronósticos. Las hojas de cálculo, junto con otros programas confeccionados a la medida, facilitan el trabajo con ambos enfoques e, incluso, con muchos otros.

Por su parte, la teoría de conjuntos ofrece una manera de clasificar los datos y sus operaciones contribuyen a modelar las interrelaciones entre los conjuntos resultantes, de modo que constituye un soporte formal para reflejar la realidad.

De esta manera se podría considerar los datos que representan el funcionamiento del negocio, sin perder generalidad, como valores cuantitativos mientras que los criterios mediante los cuales se analizan estos valores cuantitativos se dividirían en dos grupos. Con el primero se identifica el escenario en que ocurre el evento analizado y contiene los criterios complementarios o dimensiones; al segundo corresponden directamente los valoresexaminados, conocidos como criterios cuantitativos o medidas.

Desde el punto de vista de los datos en su vínculo con la realidad, los valores cuantitativos cobran sentido únicamente cuando se relacionan con los conjuntos de criterios complementarios. La obtención de información más detallada es posible gracias a la existencia de jerarquías en las dimensiones, que se manifiestan mediante el establecimiento de relaciones de pertenencia entre las subdivisiones lógicas del conjunto de datos de cada dimensión.

Hasta aquí se ha expuesto un acercamiento informal a los elementos multidimensionales. Ahora bien, resulta conveniente contar con definiciones tales que permitan expresar cada componente del modelo multidimensional independientemente de cómo se instrumente una solución posterior. Para ello varios autores han publicado estudios que difieren unos de otros [2, 3, 4]. Asimismo, se detectó que los enfoques, en sentido general, no eran suficientemente abarcadores para acatarlos como paradigmas, ya sea por haberse planteado desde un principio objetivos más restrictivos o por omisión de elementos que se han considerado esenciales en el modelo.

Dado que no existe un punto de vista único, se consideró oportuno concebir un enfoque constructivo propio para expresar formalmente el modelo y que, a su vez, contribuyera a su aplicación práctica. Vale acotar que a continuación solo incluiremos los componentes esenciales del modelo multidimensional y que la discusión detallada se puede encontrar en [16].

En primer lugar, un dominio consiste en un conjunto de miembros que no son más que valores, todos del mismo tipo. Se considera que los dominios son conjuntos abiertos de tipos de datos en el sentido que los usuarios pueden definir nuevos tipos y que todo valor corresponde a un tipo determinado, sea o no escalar [15].

Se define un nivel como la relación de asociación entre un nombre y un conjunto de valores escalares con semántica similar. Se denota por L al subconjunto de los datos multidimensionales formado por niveles de modo que a cada l∈L se le asocia un dominio o conjunto de valores —denotado por DOM(l)— que contendrá los miembros del nivel. La cardinalidad del nivel se denotará por Card(l). Formalmente, li = {'Nombre$_i$', {DOM(l_i)}} para todo l_i∈L.

Para cualquier subconjunto de niveles L' c L es posible definir al menos una relación de precedencia Rp' tal que Rp'(l_i, l_j) donde i < (>) j y l_i, l_j∈L'. Se dice entonces que li es inferior (superior) a lj según Rp' o que l_i⟨Rp'l_j (l_i⟩Rp'l_j). Asimismo, para cualesquiera dos niveles l_i, l_j∈ L' c L y l_i⟨Rp'l_j es posible definir al menos una relación de correspondencia Rc' para toda k, k = 1, … , Card(l_i,) tal que Rc'(μ^{li}k, {μ^{lj}k}) donde μ^{li}k∈ DOM(l_i) y {μ^{lj}k} c DOM(l_j).

Cada relación de precedencia Rp'∈Rp permite establecer relaciones de correspondencia entre cada par de niveles l_i, l_j siempre que l_i ⟨Rp' l_j. Por tanto, la relación formada por una Rp' y el conjunto de todas las relaciones de correspondencia entre varios niveles consecutivos constituye una jerarquía. Si se denota por J al conjunto de las jerarquías, entonces ji∈ J se puede definir como ji= {L', $Rp^{L'}$, {$Rc^{L'}$}} para todo ji∈J donde $Rp^{L'}$∈Rp y $Rc^{L'}$⊂Rc.

A partir de lo anterior, se define una dimensión como la relación entre un nombre, un conjunto de niveles y un conjunto de jerarquías

establecidas sobre tales niveles que pueden o no compartir una relación de precedencia. Luego, el conjunto de todas las dimensiones se expresa por D, de modo que D = { d = {'Nombre', L', J'} donde L' ⊂ L y J' ⊂ J}. Asimismo, se denota por LD al conjunto de todos los niveles de las dimensiones que expresan los criterios cualitativos.

Siguiendo la definición formulada se puede establecer el conjunto de medidas M como una dimensión que solo contiene un nivel a cuyo dominio pertenecen los criterios cuantitativos. Formalmente, M ⊂ D de modo que m si m ∈M entonces m = {'Medidas', L_m, J_m} donde Lm={ l1} ⊂ L yJ_m = { L_m, Rp'^{Lm}, {Rc'^{Lm}}} ⊂ J tal que Rp'^{Lm}= {} yRc'^{Lm}= {}.

Para lograr que los datos en estudio cobren sentido y expresen la realidad de manera coherente deben vincularse los valores cuantitativos y los criterios de análisis. Para ello se define una función F en cuyo dominio se encuentran los miembros de cada dimensión. La imagen de F es el conjunto de todos los hechos y se denota por H. En otras palabras, el conjunto H está formado por las tuplas de los valores cuantitativos que corresponden a cada criterio cuantitativo.

Formalmente, se tiene F:Dom(L_d)UDom(L_m)-> H para Dom(F) = Dom(L_d) U Dom(L_m) e Img(F) = H. Lo que puede expresarse también como F(L_d, L_m) = H = (valor1, …, $valor_{rn}$) donde Card(L_d) >= 0, Card(L_m) > 0 y n = Card(L_m).

Ahora bien, en términos del análisis informacional y sin perder generalidad, se plantea que existe una función de agregación, llamada genéricamente Fa, definida para cada miembro del dominio de las medidas que permite resumirlas en correspondencia con los requerimientos analíticos. En lo adelante se hará referencia a las operaciones de agregación en sentido general utilizando el símbolo 'Ω'.

Siendo C el conjunto de todos los cubos, D'⊂D un conjunto finito de dimensiones, m⊂D'≠ {} y L_d = {U_i L_i} donde i = 1, …, Card(D') y L_d⊂L, se define un cubo c_i∈C como c_i = {'$Nombre_i$', D', m, F', F'a^m tal que F({UjDom(lj)} donde j = 1, …, Card(Ld) y l_j∈L_d) → H y F'am(U_jDom(l_j)}, j = 1, …, Card(L_d) y lj∈L_d) →ΩH. Si los datos están detallados, entonces Img(F') = Img(F'am).

Concluyendo la formalización de las estructuras de datos del modelo, se tiene que una base de datos multidimensional consiste en la relación de un nombre, un conjunto de cubos construidos sobre determinadas dimensiones —que, a su vez, están construidas sobre determinado conjunto de niveles— y un conjunto de dimensiones que aún no están asociadas a ningún cubo. Por tanto, siendo B el conjunto de todas las bases de datos multidimensionales, se tiene que:

B = {bi = {'$Nombre_i$', C', D'}, donde C' ⊂ C y D' ⊂ D para todo b_i∈ B}.

Si se combinan las relaciones de las jerarquías de cada dimensión con las funciones del cubo es posible obtener las operaciones de navegación del cubo, conocidas como drill-down y roll-up.

Finalmente, el modelo de restricciones del enfoque relacional garantiza la integridad, la consistencia y la veracidad de los datos. Estas restricciones son válidas para el enfoque dimensional, aunque sus estructuras de datos y las operaciones sean diferentes. Además, las definiciones de las propias estructuras ofrecen restricciones del modelo. Por ejemplo, no pudiera existir una dimensión sin un nivel asociado, ni un cubo sin al menos una medida y una dimensión.

Como en todo fenómeno existen reglas del negocio que no se pueden expresar mediante las restricciones propias del modelo o los predicados, en cuyo caso será la programación la que permita controlar el comportamiento de los datos para cumplir con los requerimientos de los usuarios finales.

Expresar la realidad en términos analíticos pudiera resultar complejo, sin embargo, al examinar conceptual y prácticamente las estructuras dimensionales se evidencia la riqueza del modelo para reflejar el comportamiento de los principales indicadores de una organización en correspondencia con sus escenarios y su ejercicio en el marco de las soluciones de inteligencia de negocios.

5. AMBIENTE PARA LA INSTRUMENTACIÓN DEL DATA WAREHOUSING

Resulta trascendental la veracidad del data warehouse como reflejo del funcionamiento y de las características propias de la empresa y,para lograrlo, es esencial que la representación física

del DW constituya una proyección concreta del modelo o diseño lógico de la empresa.

A partir de las especificaciones de la arquitectura de datos de tres capas y de la anterior aseveración se hace necesario separar el diseño físico de las capas de datos reconciliados y de datos derivados manteniendo la interrelación de dependencia entre ellas y el modelo de datos de la empresa. Se trata entonces de ofrecer un ambiente computacional con un conjunto de herramientas genéricas que comprendan el diseño físico de estas dos capas de la arquitectura y sus respectivas etapas de población que, interactuando entre sí, conformen un proyecto que facilite el diseño y la puesta a punto del data warehouse de la organización.

La complejidad de la solución se manifiesta, por un lado, en el suministro de facilidades para satisfacer los requerimientos de instrumentación y explotación del proceso de data warehousing en función no solo de las peculiaridades de la organización, sino también de la arquitectura conceptual de los datos y, por otro, en la contribución a perfeccionar las funcionalidades delambiente y habilitarlo en nuevos entornos de acuerdo con las exigencias de los diseñadores del data warehouse.

Para implementar el modelo de solución planteado nos hemos apoyado en las fortalezas de la programación orientada a objetos, en particular en la herencia y el polimorfismo. El ambiente para la instrumentación del data warehousing fue desarrollado en .NET creando un prototipo para Microsoft SQL Server. En la figura 2 puede observarse el flujo de ambos procesos de población instrumentados sobre sendas herramientas genéricas.

Las herramientas genéricas tienen como objetivo facilitar la población del DW en su conjunto persiguiendo que el ambiente se pueda extender a diferentes gestores de bases de datos tanto relacionales como dimensionales. De la misma manera se desea que puedan ser incluidos nuevos tipos de transformaciones en el proceso de población del data warehouse empresarial.

Como respuesta a estos propósitos, en el caso del DWE la arquitectura basada en extensiones constituye el fundamento conceptual y práctico de la herramienta genérica que responde al procedimiento de diseño del proceso ETCL así como a la realización del proceso de población del DWE en el marco de la creación y la explotación de un data warehouse. Asimismo, sirve de base a la aplicación del modelo de solución en función de diversas plataformas al igual que al perfeccionamiento de la herramienta, mediante la inclusión de nuevas funcionalidades.

La figura 3 muestra las relaciones que se establecen entre los componentes esenciales del procedimiento de diseño del proceso de población del DWE y la arquitectura basada en extensiones que la sustenta. Un marco denotado con líneas discontínuas engloba las acciones del flujo ETCL como componente y el otro marco agrupa el resto de las componentes del procedimiento, mientras que las flechas señalan la dirección de la interacción entre los elementos del modelo. Por otra parte, las imágenes cilíndricas denotan, como de costumbre, el soporte físico y la relación con uno o más soportes en dependencia del tipo de extensión.

Cada grupo de extensiones cumple un objetivo específico dentro del diseño del proceso de población.

- **Extensiones de extracción:** provee una alternativa a la dependencia del soporte físico y la diversidad de orígenes de los datos.

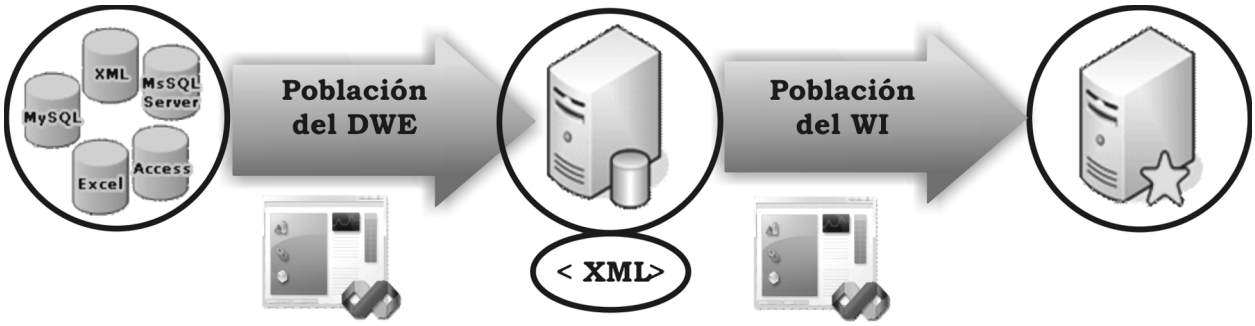

Fig. 2. *Población del data warehouse.*

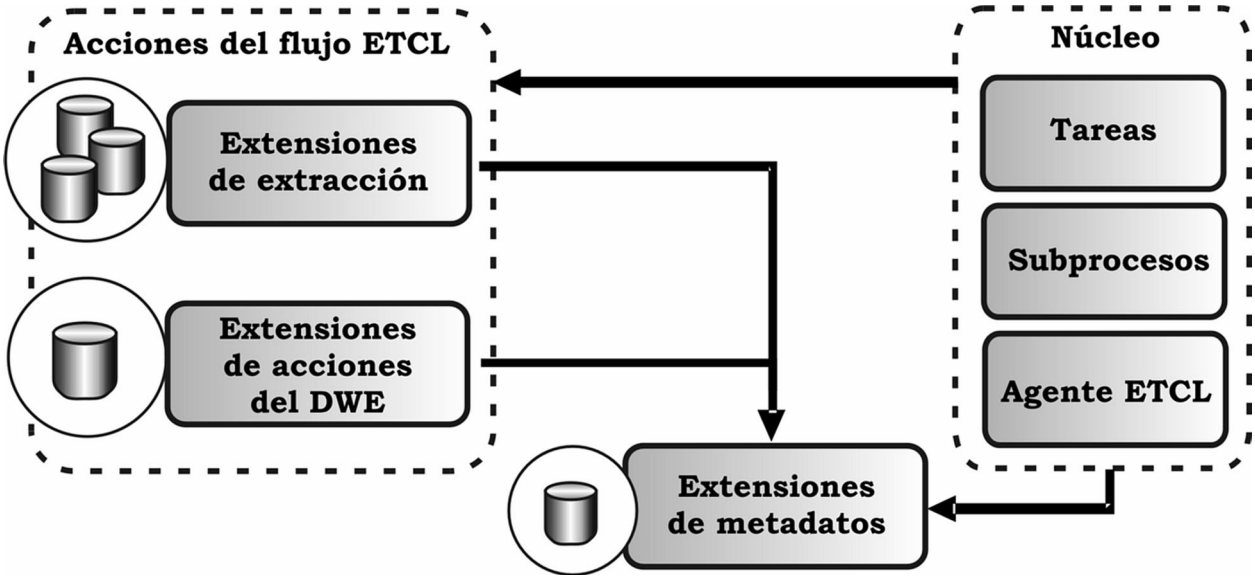

Fig. 3. *Modelo de solución para la población del DWE.*

- **Extensiones de acciones del DWE:** responde a la dependencia y la variabilidad del soporte físico del data warehouse empresarial.
- **Extensiones de metadatos:** proporciona una alternativa a la dependencia y la variabilidad del soporte físico de los metadatos.

Los grupos de extensiones se instrumentaron a partir de los patrones de diseño de la programación orientada a objetos. Por ejemplo, en la jerarquía de acciones existe una clase abstracta para cada tipo de transformación y chequeo de calidad, lo cual garantiza el dinamismo de las acciones en el flujo ETCL aplicando el Patrón de Estrategia. De igual manera se añade en la jerarquía una clase abstracta que sirve de enlace entre los componentes del núcleo y el grupo de acciones incluidas en una implementación concreta de la extensión de acciones del DWE y para un formato físico específico. En este mismo sentido se implementaron el resto de los grupos de extensiones cada uno representado por un patrón de diseño diferente en relación con sus objetivos.

El proyecto dirigido a la población del DWE está conformado por diferentes módulos. Teniendo en cuenta el papel que desempeña cada uno en su relación con el modelo de solución, los módulos se clasifican en tres grupos:

- **Grupo de gestión:** proporciona los medios para la instrumentación y el control del proceso de población de un DWE por parte de los diseñadores y los administradores e incluye los módulos con los que interactúan directamente estos usuarios.
- **Grupo base:** constituye los cimientos de la funcionalidad e implementación ulterior de las extensiones, dado que en sus módulos se incluyen las clases bases de las jerarquías de los tipos de extensiones.
- **Grupo de extensiones:** está compuesto por cada una de las implementaciones concretas de las clases abstractas.

En la figura 4 se resumen los grupos descritos y se expresa el vínculo existente entre ellos y los distintos roles de los usuarios de la aplicación.

Dada la importancia de la extensibilidad lograda a partir del modelo de solución propuesto se sugiere que se continúe implementando extensiones que respondan a nuevas acciones del proceso de Extracción, Transformación, Depuración y Carga, con el propósito de enriquecer el ambiente de población.

Por otra parte, en el ámbito matemático-computacional las estructuras dimensionales formalizadas pueden expresarse en términos de abstracciones que resultan cómodas y claras, con vistas a generar un modelo de clases sobre el cual se fundamenta la concepción de la herramienta genérica que permite implementar y enriquecer el proceso de población del data warehouse en diversas plataformas desde la perspectiva informacional.

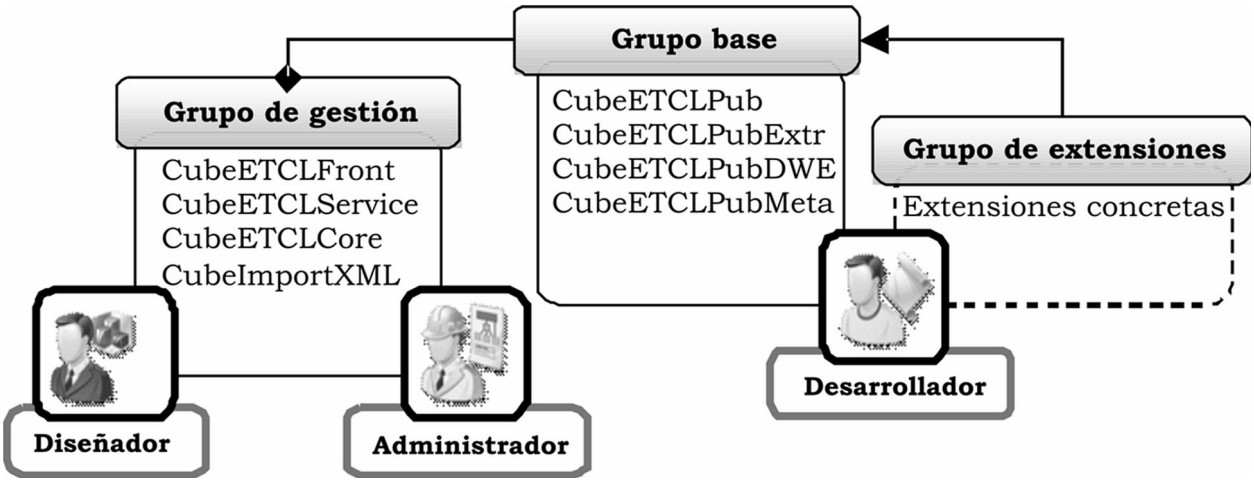

Fig. 4. *Instrumentación de la herramienta genérica del DWE.*

El modelo de clases concebido que resume una base multidimensional consiste en:

- Base de Datos
  - o Nombre
  - o Lista de Dimensiones
  - o Lista de Cubos
- Cubo
  - o Nombre
  - o Lista de Dimensiones
- Dimensión
  - o Nombre
  - o Lista de Niveles
- Nivel
- Nombre
- Lista de Cubos
- Lista de Dimensiones
- Lista de Niveles

El modelo de solución para la instrumentación del proceso de población del WI comprende un conjunto de interfaces que solo exponen qué características y operaciones corresponden a cada tipo de estructura dimensional y un conjunto de clases que implementan soluciones particulares para las plataformas específicas apoyadas en la herencia.

La aplicación consta de varios proyectos. En la figura 5 se distingue la funcionalidad de cada proyecto según sus objetivos, así como su integración en la solución propuesta. El proyecto CubeDesign, instrumenta la creación del WI para lo cual se apoya en el resto de los ensamblados mediante una interfaz de usuario dirigida al creador/administrador del WI. Se ha subrayado el proyecto CubeDSOWrapper por simbolizar el prototipo desarrollado para verificar la validez de la solución propuesta.

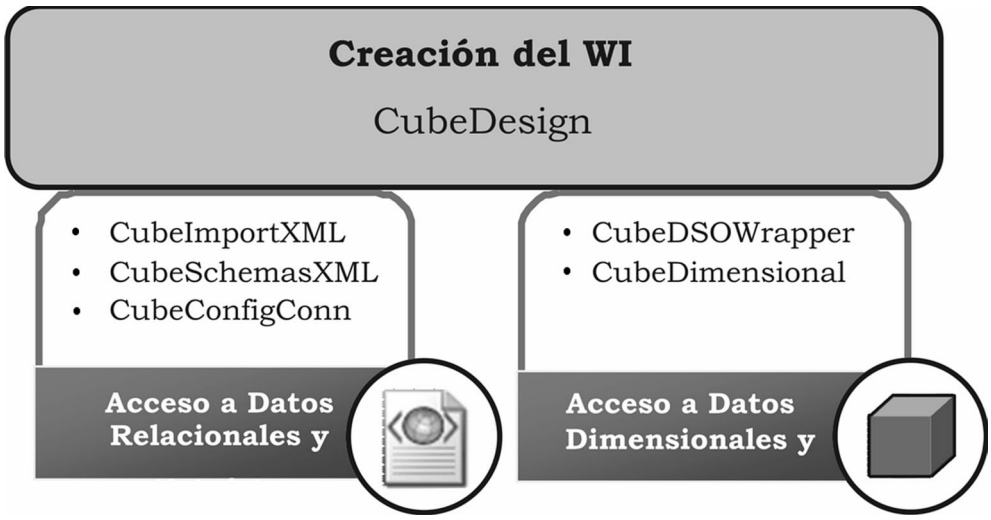

Fig. 5. *Instrumentación de la herramienta genérica del WI.*

La herramienta para la población del WI fue diseñada bajo los mismos principios de la herencia para lograr la extensibilidad deseada y el aislamiento del formato físico de las estructuras dimensionales. Basándonos en el modelo de clases se creó un conjunto de interfaces dentro del proyecto CubeDimensional que representan cada una de las estructuras dimensionales y se implementó el prototipo heredando de dichas interfaces.

Teniendo en cuenta que uno de los objetivos de los DW consiste en el aislamiento de los propósitos informacionales de los fines operacionales, sería aconsejable continuar trabajando en la formalización del modelo multidimensional y su instrumentación genérica, en particular, en el desarrollo de un álgebra que permita expresar las restricciones y operaciones del modelo de datos.

Por último, la relación intrínseca de dependencia entre las dos etapas de población dictamina la generación de los conjuntos de datos y los metadatos necesarios para el procesamiento en el WI durante la población del DWE. Por otra parte, existe una particularidad de la integración que fomenta la necesidad de sincronización entre ambos procesos.

El ambiente para la instrumentación del data warehousing proporciona alternativas para determinar las tareas que conforman los subprocesos y las dependencias entre ellas, ya sea como componentes del proceso de población del data warehouse empresarial en su papel estructural y funcional, o como ejecutor del procesamiento multidimensional en su carácter funcional.

6. CONCLUSIONES

Penetrar en el data warehousingen la arista teórico-conceptual al incursionar en una formalización matemáticadel modelo dimensional, en el ángulo metodológico al profundizar en el procedimiento de diseño del proceso de población y en el ámbito aplicativo al servir de fundamento para la concepción de un ambiente que permite implementar y enriquecer el proceso de población del data warehouse en diversas plataformas constituyen los resultados más significativos.

El ambiente para la instrumentación del data warehousing fundamenta la aplicación de la estrategia de desarrollo en espiral con vistas al crecimiento y el perfeccionamiento del data warehouse, de manera que conserve su carácter de pilar esencial para el tratamiento de datos estructurados orientado a la toma de decisiones en cualquier organización o empresa.

El modelo conceptual y las soluciones arquitectónicas en su perfilutilitario, adaptativo y progresivo pretenden contribuir a las exigencias evolutivas de las soluciones de inteligencia de negocios desde la perspectiva de los datos.

Quedan abiertas aún numerosas ramas de investigación en el marco de la Inteligencia de Negocios, se pudiera intentar complementar los metadatos desde el punto de vista semántico. Asimismo, el empleo de métodos de minería de datos, la aplicación de cuadros de mando integrales, el aprovechamiento del ambiente Web, el tratamiento de datos no estructurados, entre otros, pudieran propiciar el modo de facilitarle alos especialistas y ejecutivosde una organización los diversos tipos de información que requieren para tomar decisiones pertinentes, oportunas y certeras.

BIBLIOGRAFÍA

1. Álvarez Sánchez, R. Estudio teórico y conceptual sobre el proceso de población del Data Warehouse Empresarial. Tesis de Licenciatura en Ciencia de la Computación, dirigida por la Dra. Lucina García Hernández y la Lic. Lis Velázquez Vidal. Facultad de Matemática y Computación. Universidad de La Habana, Cuba, 2002.
2. Cabibbo, L. Torlone, R.A LogicalApproachto Multidimensional Databases, Sixth International Conference on Extending Database Technology, Universidad de Roma, Italia, 1998.
3. Carpani, Fernando. CMDM: Un Modelo Conceptual para la Especificación de Bases Multidimensionales. Tesis de Maestría, Instituto de Computación-Facultad de Ingeniería. Universidad de la República. Pedeciba Informática, Montevideo, Uruguay, 2000.
4. Cid Díaz, Hedrie. Almacén de Datos: Propuesta de Formalización del Modelo Dimensional. Navegando y Construyendo Agregaciones. Tesis de Licenciatura en Ciencia de la Computación, dirigida por el Lic. Alfredo Somoza Moreno. Universidad de La Habana, Cuba, 2002.
5. Devlin, B. Data Warehouse from Architecture to Implementation. Addison Wesley Longman, Inc. 1997.

6. Ferrá Díaz, R. Preparación de datos para la extracción de conocimiento en un Data Warehouse. Tesis de Licenciatura en Ciencia de la Computación, dirigida por la Dra. Lucina García Hernández. Facultad de Matemática y Computación, Universidad de La Habana, Cuba, 2007.
7. García Hernández, L.; Oliva Santos, R.; Prendes Arencibia, H.; Velázquez Vidal, L; Veliz Monteagudo, M. La inteligencia de negocios desde la perspectiva de los datos. 4to. Evento Nacional de Informáticos. COPEXTEL, Cuba, 2008.
8. Inmon, W. H. www.cait.wustl.edu/cait/papers/prism Prism Solutions, Inc. 1995.
9. Inmon, W. H. Building the Data Warehouse. Wiley Computer Publishing, 2002.
10. Kimball, R. y Ross, M. *The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling.* Wiley Computer Publishing, 2002.
11. Kimball, Ralph y Caserta, Joe "The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data". Wiley Computer Publishing. 2004.
12. Kimball, Ralph "A Dimensional Modeling Manifesto". DBMS Online. Agosto de 1997.
13. Magic Quadrant for Business. Intelligence Plat forms, IQ07, 2007.
14. Olsen, Jack E. "Data Quality: The Accuracy Dimension". Morgan KaufmannPublishers. 2003.
15. Pendse, Nigel. What is OLAP? The OLAP Report. www.olapreport.com. 2003.
16. Velázquez Vidal, L. Herramienta genérica para la población del Warehouse Informacional. Tesis de Maestría en Ciencia de la Computación, dirigida por la Dra. Lucina García Hernández. Universidad de La Habana, Cuba, 2009.
17. Velázquez Vidal, L. y Veliz Monteagudo, M. Estudio teórico y conceptual sobre Data Warehouse. Tesis de Licenciatura en Ciencia de la Computación, dirigida por la Dra. Lucina García Hernández. Universidad de La Habana, Cuba, 2000.
18. Veliz Monteagudo, Mijail. Herramienta genérica para la población del Data Warehouse Empresarial. Tesis de Maestría en Ciencia de la Computación, dirigida por Dra. Lucina García Hernández. Universidad de La Habana, Cuba, 2009.
19. White, C. Intelligent Business Strategies: OLAP in the Database. Columna publicada en la revista electrónica. *DM Review Magazine*, Junio 2003.