Universidad de las Ciencias Informáticas

Facultad 4

Título: GESTCON MART, DATA MART PARA LA GESTIÓN DEL CONOCIMIENTO

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autores: Michel Diaz Llerena

Yanetsi Mestre Morales

Tutores: Lic. Eddy Manuel Infante

Ing. Julio Cesar Díaz Vera

Ciudad de La Habana, Cuba

Julio, 2007

No basta con alcanzar la sabiduría,

es necesario saber utilizarla.

Marco Tulio Cicerón

DECLARACIÓN DE AUTORÍA

Declaramos que somos los únicos autores de este trabajo y autorizamos a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmamos la presente a los ____ días del mes de _____ del año _____.

_____	_____
Yanetsi Mestre Morales	Michel Diaz Llerena
Autor	Autor
_____	_____
Lic. Eddy Manuel Infante	Ing. Julio Cesar Díaz Vera
Tutor	Tutor

Agradecimientos de Yanetsi:

A mi adorada madre, que siempre me incitó a luchar por el futuro, por todo ese amor sin límites, por su apoyo, dedicación, sacrificio y por ser guía y ejemplo para mí.

A mi querida abuelita Elia, por quererme tanto, por la confianza depositada en mí, por darme ánimo en todo momento y por sus sabios consejos.

A mi queridísima tía Estrella, por su aliento, su cariño, su preocupación, su apoyo y por creer en mí.

A Rolando, por ser el padre que nunca tuve.

A mi hermanito Roli, por alegrar mi vida.

Al resto de mi familia, por confiar en mí, por estar al tanto y darme ánimo.

A todos mis profesores, por su ayuda y contribuir en mi formación profesional.

Agradecimientos de Michel:

A mi abuela adorada Digna, por confiar en mí, por su dedicación, su amor, su aliento y sus consejos.

A mis queridos padres, por su apoyo, cariño, ayuda, ánimo, confianza y preocupación tanto tiempo.

A mi tío Claro, por ayudarme en los estudios y velar por mí.

A todos los miembros de mi familia, por darme ánimo en todo momento.

A los profesores que han contribuido con sus conocimientos y experiencia en mi formación profesional, pero en especial a Lili y Aidé.

En general queremos agradecer a nuestro tutor Julio, por ser tan atento, por su ayuda y su empeño en la realización de este trabajo.

A nuestros amigos y compañeros de aula.

A todos los que de una forma u otra han tenido que ver con la realización de la tesis.

A la Revolución y a Fidel.

A nuestros padres y familias

Resumen

La gestión del conocimiento es un recurso vital para el desarrollo de las organizaciones y un pilar fundamental para aumentar la productividad y el crecimiento económico. Sin embargo en los proyectos productivos de la Universidad de las Ciencias Informáticas (UCI), la gestión del conocimiento no se maneja, para ello en el presente trabajo se lleva a cabo el diseño y la implementación de un Data Mart para esa área particular de la Universidad, el mismo servirá de soporte a una base de experiencia para la gestión de proyectos de software, con el propósito de hacer un mejor uso de la información y facilitar su posterior análisis en el proceso de toma de decisiones, al hacerlas más rápidas y exactas.

Para realizar el diseño del Data Mart se utilizó la metodología *Data Warehouse Engineering Process (DWEP),* la cual está basada en RUP y UML, lenguaje de modelado visual que se usa para especificar, visualizar, construir y documentar artefactos de un sistema de software.

PALABRAS CLAVE

Data Warehouse, Data Mart, gestión del conocimiento.

Índice

1 Capítulo 1: Introducción

1.1 Introducción

La nueva era de la economía del conocimiento ha jugado un papel sorprendentemente creciente en el desarrollo científico, tecnológico y económico, donde el factor esencial de progreso es el conocimiento. Esta nueva sociedad, con organizaciones basadas en el aprendizaje, tiene como capital más preciado al ser humano y como elemento básico la capacidad de interpretar creativamente la información disponible, además ofrece una multitud de oportunidades, entre las cuales se encuentra el desarrollo del software para una mejor gestión del conocimiento.

Aunque son muchas y variadas las definiciones existentes de Gestión del Conocimiento se puede definir como el proceso sistemático de buscar, organizar, filtrar y presentar la información, con el objetivo de mejorar la comprensión de las personas en un área específica de interés. Una adecuada gestión del conocimiento garantizará el éxito de las organizaciones dentro de un mercado tan competitivo como el actual. Es por esta razón que la gestión del conocimiento se torna fundamental "…así como la ingeniería mecánica y electrónica ofrecen teorías, métodos y técnicas para la construcción de automóviles, la ingeniería del conocimiento nos equipa con metodología científica para analizar y generar conocimiento"(SCHREIBER *et al.* 2000).

La gestión del conocimiento debe convertirse en una disciplina práctica que ayude a mejorar la gestión interna de las organizaciones y que propicie el desarrollo de una cultura organizacional, donde la integración e interacción de la información y el conocimiento no tengan barreras.

En el caso de las empresas cubanas la introducción de tecnologías de la información y las comunicaciones en el manejo del conocimiento tiene como objetivo fundamental el uso más racional de los recursos, el logro de una mayor productividad y la obtención de los productos con una mayor calidad.

Se debe tener en cuenta que el uso indistinto de los términos datos, información y conocimiento es erróneo. Una primera aproximación de sus diferencias es la siguiente: los datos están localizados en el mundo y el conocimiento está localizado en agentes (personas, organizaciones...) mientras que la información adopta un papel mediador entre ambos conceptos.

Los datos son observaciones, hechos o imágenes que una vez formalizados, contextualizados, filtrados y resumidos, constituyen información, finalmente, dicha información enriquecida por ideas, procedimientos y reglas que permiten realizar acciones y tomar decisiones forma el conocimiento (LIEBOWITZ 1999).

En definitiva el conocimiento existe dentro de las personas y se deriva de la información y esta a su vez está compuesta por datos que han recibido un procesamiento y tienen un significado claro y definido. El conocimiento, sin embargo, implica generar acción con la información que proviene de esos datos.

En la economía de los países a escala mundial, la gestión de la información ocupa cada vez más, un espacio mayor y reconocen a la información como un recurso indispensable para ampliar su competitividad, aumentar la calidad y la satisfacción de los clientes. Su correcta gestión es una herramienta fundamental para la toma de decisiones, la formación del personal, la evaluación de los productos, la determinación de los errores y el control de los procesos; es un recurso vital para el desarrollo de la organización, pero sólo se convierte en conocimiento cuando los individuos la aplican para la resolución de un problema. Sin una adecuada gestión de la información, es imposible llegar a la gestión del conocimiento.

Hoy en día existen diversos tipos de sistemas de soporte para la toma de decisiones, pero el que ha tenido más auge a escala mundial en las grandes instituciones sin duda ha sido el Data Warehouse (DWH) o almacenes de datos, convirtiéndose en el centro de atención de las organizaciones, puesto que provee un ambiente para hacer un mejor uso de la información administrada por diversas aplicaciones operacionales.

Las aplicaciones para soporte de decisiones basadas en un data warehouse, pueden hacer más práctica y fácil la explotación de los datos para una mayor eficacia del negocio, al brindar la información en el nivel de detalle adecuado para posteriores análisis, proporcionándole a la empresa una enorme ventaja competitiva.

1.2 Problema Científico

La Universidad de las Ciencias Informáticas se adentra en el mundo de la gestión del conocimiento al plantearse elevar la calidad a niveles de excelencia en el proceso de producción de software, ya que como

se ha visto anteriormente, el conocimiento constituye un pilar fundamental para aumentar la productividad y el crecimiento económico.

Debido a que el volumen de información referente a los problemas y soluciones presentes en la realización de diversas tareas en los proyectos productivos de la Universidad, no se maneja, ni se cuenta con ningún sistema para realizar dicho trabajo, es necesario contar con herramientas software para la gestión del conocimiento, permitiendo la diseminación, utilización y traspaso de experiencias e información entre todos sus integrantes, facilitando la resolución de problemas al disponer de mecanismos que utilizan información almacenada sobre situaciones anteriores, garantizando con ello que la Universidad aprenda todos los días, logrando una mayor eficiencia y efectividad en la producción.

Puesto que las bases de datos relacionales requieren que los usuarios sean expertos en lenguajes de consulta y no permiten la presentación de los datos en forma adecuada para las tareas de análisis, sumado a la necesidad de contar con datos estructurados sobre los que se puedan ejecutar consultas complejas, obtener reportes y realizar análisis con facilidad y fiabilidad en un tiempo de ejecución aceptable conlleva a la utilización de Data Warehouses, para que las aplicaciones de apoyo a la gestión del conocimiento utilicen dicho almacén de datos en la extracción de la información a fin de facilitarles el trabajo a los responsables de la toma de decisiones en los proyectos productivos.

Por tanto, la carencia de un Data Warehouse o al menos de un Data Mart (DM) sobre el área de conocimiento para la gestión de proyectos en la UCI, limita las oportunidades de desarrollo de software debido a la repetición de errores de un proyecto a otro, perdiéndose con el tiempo las experiencias positivas en la implantación de mejoras a los procesos de gestión de proyectos y reduciéndose las probabilidades de aprender de los ya desarrollados de modo que se puedan extender sus éxitos a los nuevos proyectos. Por consiguiente, se seleccionará el siguiente problema científico ¿Cómo garantizar la reutilización del conocimiento generado en la gestión de proyectos productivos en la UCI a partir de la implementación de un Data Mart?

1.3 Objeto de estudio, objetivos e hipótesis

Teniendo en cuenta lo anterior se define como objeto de estudio de la investigación las metodologías de diseño e implementación de un Data Warehouse en la UCI, así como objetivo general diseñar e implementar un Data Mart, que brinde la posibilidad de analizar los problemas y las soluciones en

diferentes proyectos productivos y ponerlo a disposición de sus líderes de manera que pueda facilitarles el trabajo. Se puede decir además que el campo de acción está centrado en el proceso de implementación de un Data Mart, nombrado GestCon Mart.

Para dar cumplimiento a los objetivos anteriormente planteados se definen las siguientes tareas:

- Hacer un estudio preliminar del problema y la situación actual.
- Investigar y analizar las técnicas de diseño de los Data Warehouses.
- Revisar y seleccionar los gestores de Base de Datos que dan soporte a los Data Warehouses.
- Buscar y analizar la bibliografía referente a los Data Warehouses.
- Establecer un diagnóstico de las tendencias actuales y tomar una posición al respecto.
- Obtener los requisitos del sistema.
- Hacer un análisis y discutir la arquitectura del Data Mart.
- Obtener el modelo de datos más idóneo para la implementación del GestCon Mart.
- Implementar el GestCon Mart para el almacenamiento de los datos referente a los proyectos productivos de la Universidad de Ciencias Informáticas.

Hipotéticamente se plantea que si se dispone de un Data Mart con datos almacenados sobre la gestión de proyectos informáticos, entonces sus líderes serán capaces de tomar mejores decisiones a nivel estratégico por contar con una herramienta que les permita reutilizar de manera eficiente las experiencias previas tanto propias como de otros líderes

VI: Disposición de un Data Mart.

VD: Toma de mejores decisiones a nivel estratégico.

1.4 Organización del documento

El presente documento se estructura en resumen, tres capítulos de contenidos, conclusiones, recomendaciones, referencias bibliográficas y anexos.

En el siguiente capítulo se recogen conceptos, objetivos y características de los Data Warehouses, así como las metodologías, tecnologías y software empleados en el diseño, además se realiza un análisis de las diferencias y funcionalidades de los principales gestores de bases de datos que dan soporte a los data Warehouses.

En el capítulo 3 se lleva a cabo la implementación del GestCon Mart, se presenta una breve descripción del mismo, el proceso de diseño de los DWH así como la aplicación de la metodología seleccionada para realizar su diseño a través de tres niveles: conceptual, lógico y físico.

2 Capítulo 2: Características del sistema

2.1 Introducción

En el presente capítulo se hace un análisis de conceptos, objetivos y características de los Data Warehouses, así como de las principales tecnologías, software y tendencias que se están empleando en el mundo para el diseño del mismo. Se analizan también las diferencias y funcionalidades de los principales gestores de bases de datos, para la exploración de datos y la extracción de conocimiento, además de fundamentar cuáles de las tecnologías se utilizarán para el desarrollo de este trabajo.

2.2 Data Warehouse

En las últimas décadas se ha venido utilizando los Data Warehouses para facilitar el análisis y comprensión de los datos de las empresas. Las dos personalidades más destacadas en este tema sin duda son William H. Inmon conocido como el padre de los Data Warehouses y Ralph Kimball considerado el principal promotor del enfoque dimensional para el diseño de almacenes de datos, ambos han desarrollado sus propios enfoques, modelos y arquitecturas de DWH.

Existen varios criterios y definiciones sobre qué es un Data warehouse, a grandes rasgos se puede decir que es una base de datos que almacena información proveniente de diferentes bases de datos operacionales y/o externas, con el objetivo de consolidar dicha información y hacerla disponible para que los usuarios finales puedan fácilmente ejecutar consultas, confeccionar reportes y realizar análisis para la toma de decisiones.

Una definición más formal es la brindada por William H. Inmon en (INMON, WILLIAM. H. 1992) donde plantea que "… es una colección de datos orientado a temas, integrado, no volátil y de tiempo variante, organizados para dar soporte al proceso de ayuda a la toma de decisiones", puesto que permiten analizar la información consolidada según diferentes puntos de vista.

2.2.1 Objetivos y Características de un Data Warehouse

En un data warehouse los objetivos más importantes son:

- Proveer una única visión de los clientes a través de toda la compañía.
- Proveer la mayor cantidad de información a la mayor cantidad de personas dentro de la organización.
- Mejorar el tiempo de emisión de informes.
- Monitorear el comportamiento de los clientes.
- Mejorar la capacidad de respuesta a las cuestiones del negocio.
- Mejorar la productividad.

Tal como se ha definido por Inmon, los datos de un data warehouse tienen una serie de características, a continuación se explican cada una de ellas.

Integrada: Los datos almacenados en el data warehouse deben integrarse en una estructura consistente, por lo que las inconsistencias presentes entre los diversos sistemas operacionales deben ser eliminadas.

Orientada a temas: Los datos son organizados por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. Por ejemplo, todos los datos sobre personas pueden ser consolidados en una única tabla del data warehouse. De esta forma, las peticiones de información sobre personas serán más fáciles de responder dado que toda la información reside en el mismo lugar.

Variante en el tiempo: Los datos no se actualizan, sino que se almacena el historial de ellos, es decir, el conjunto de valores que el dato ha tenido a lo largo de su historia, asociando a cada dato una referencia de tiempo, con el fin de poder identificar los distintos valores que dicho dato ha ido tomando a lo largo de su ciclo de vida.

No volátil: El almacén de información de un data warehouse existe para ser leído, y no modificado. La información es por tanto permanente, una vez cargado los datos en el DW, deben mantenerse invariables, permitiéndose sólo realizar dos tipos de operaciones: la carga inicial de datos y el acceso a los mismos. No existe el proceso de actualización de datos, sólo de carga.

2.2.2 Métodos más usados en la construcción de Data warehouses

Los modelos propuestos por William H. Inmon y Ralph Kimball para llevar a cabo el diseño de un DWH son los más aplicados en la actualidad, coincidiendo en que un Data Mart o un data warehouse independiente no satisface las necesidades que tienen las compañías a escala corporativa de acceder inmediatamente y con facilidad a sus datos, pero sus criterios difieren en cuanto al modelo de datos y a las arquitecturas.

El término Data Mart es usado para designar a los almacenes de datos cuyo ámbito es más reducido, normalmente un departamento o área específica dentro de la empresa, es definido por Ralph Kimball como bodegas de datos con información de interés particular para un determinado sector de la empresa y aunque su enfoque sea para una sola perspectiva departamental, no lo exime de tener que seguir los lineamientos generales de implementación que posee el Data warehouse (KIMBALL 1996).

Ralph Kimball propone como modelo de datos al modelo dimensional, el más popular en las soluciones que se implementan de manera práctica, el cual facilita a los usuarios finales las consultas y el análisis. Se caracteriza por ser sencillo de crear, extremadamente estable en presencia de cambios, además de mostrarse muy intuitivo y comprensible; el autor sugiere el uso de este modelo de datos para el desarrollo de los Data Marts y del Data Warehouse (KIMBALL and CASERTA 2004). También William H. Inmon reconoce al modelo dimensional como el mejor para el desarrollo de los Data Marts por las ventajas brindadas, pero propone la construcción del Data Warehouse basado en el modelo entidad relación. La idea de Inmon se basa en que el modelo entidad relación es mucho más rico y adaptable que el dimensional (INMON, WILLIAM 2002).

En cuanto a la arquitectura William H. Inmon en su libro “Building the Data Warehouse” (INMON, WILLIAN H. 2005) plantea que la construcción del Data Warehouse no debe ser sustituida por la implementación de varios Data Marts. Resaltando que la excusa para no desarrollar un almacén de datos la mayoría de las veces es por no contar con un gran presupuesto, la sustitución de este por los Data Marts trae desventajas puesto que están diseñados para un área particular de la empresa, lo que trae consigo diferencias entre las estructuras de datos de los mismos, que al integrarlos en el Data Warehouse algunos no serán reusables, ni flexibles, ni útiles para la reconciliación que se necesita. Inmon manifiesta que el proceso de construcción del Data Warehouse parte de los sistemas operacionales existentes, creándose

áreas de diferentes temas, cuando existan una cierta cantidad de estas, el Data Warehouse inicia el proceso de población de las áreas de una manera integrada, una vez concluido se comienza a dar respuestas a las inquietudes de los usuarios; empezando así el florecimiento del nivel departamental a medida que se tienen más datos en el Data Warehouse, y es en este punto del desarrollo cuando se centra la atención en las cuestiones de los diferentes departamentos, para definir y crear los Data Warehouse departamentales, los Data Marts.

Ralph Kimball en desacuerdo con la arquitectura propuesta por William H. Inmon resalta en su libro "The Data Warehouse ETL Toolkit, Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data" (KIMBALL and CASERTA 2004), que los Data Marts están basados en los datos de la fuente y no en la visión departamental, en otras palabras que el Data Mart es sólo una parte de un producto orientado a la compañía, los cuales deben consistir en una continua pirámide de estructuras dimensionales idénticas, comenzando siempre con los datos atómicos. Plantea también que la idea de construcción de un Data Warehouse centralizado no es realista, siendo más real construirlo en un ambiente descentralizado e incremental, porque las empresas están en constante cambio, adquiriendo nuevas fuentes de datos y necesitando nuevas perspectivas, propone además, centrarse en trazar estrategias adaptables e incrementales basándose en una idealista visión de controlar toda la información antes de construir el Data Warehouse. Por esta razón manifiesta que el proceso de construcción de un almacén de datos parte de los sistemas operacionales existentes, creando los diferentes Data Marts basados en la información de dichas fuentes, para luego de tenerlos desarrollados y funcionales se comience con la construcción del Data Warehouse basado en la información que éstos contienen.

En la actualidad este método es el más usado, gracias a las diferentes ventajas que proporciona, permitiendo a las empresas acometer los proyectos de manera separada y de esta forma reducir los efectos negativos que tendría fracasar en un intento por construir un Data Warehouse.

2.3 Arquitectura conceptual de los datos

La estructura que reúne todos los componentes de un Data Warehouse es conocida como arquitectura (PONNIAH 2001), esta es la forma de representar la organización total de datos, comunicación, procesamiento y presentación. Como dice Paulraj Ponía"…la arquitectura incluye todo lo que se necesita para preparar y guardar los datos. Por otro lado, también contiene todos los recursos para distribuir la información desde el data warehouse. Está compuesta más allá de reglas, procedimientos y funciones que

permiten al almacén de datos trabajar y cumplir los requisitos de la empresa… Define las normas, medidas, diseño general, y técnicas de apoyo. ’’(PONNIAH 2001)

Para resolver los problemas en la actualidad se hace uso de diferentes arquitecturas conceptuales de datos que se definen a nivel lógico. En un Data Warehouse suele representarse varias capas a través de las cuales circulan los datos, de modo que estos se obtienen a partir de los de la capa previa (JARKE *et al.* 2003), nombrándose de acuerdo al número de capas que abarcan.

Como se expone en (VIDAL and MONTEAGUDO 2000), en la arquitectura de Datos de **Una Sola capa** la información se guarda sólo una vez en el Data Warehouse, almacenándose únicamente los datos de tiempo real, sobre los cuales actúan sistemas informacionales y operacionales, esto puede traer disputas ya que ambos sistemas actúan sobre el mismo conjunto de datos y quizás en el momento en que se precisen no estén disponibles para los fines operacionales porque pueden estar siendo consultados y mientras esto sucede no es posible realizar actualizaciones. Figura 2.1

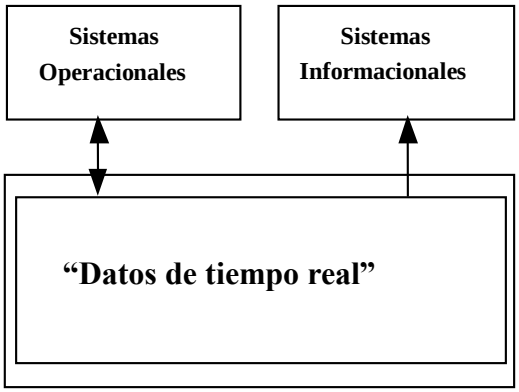


Figura 2.1 Arquitectura de datos de una sola capa

En la arquitectura de Datos de **Dos Capas** se perfecciona lo antes expuesto, conteniendo además de una capa inferior, donde se contemplarán los datos de tiempo real utilizados por las aplicaciones operacionales en modo lectura/escritura, una superior, para el almacenamiento de los datos derivados utilizados por las aplicaciones informacionales, éstos pueden ser una copia directa de los datos de tiempo real o pueden obtenerse mediante la aplicación de un algoritmo, aumentando de esta forma los requerimientos de almacenamiento debido a la duplicación de información pero garantizando el acceso de los sistemas operacionales en cualquier instante de tiempo. Figura 2.2

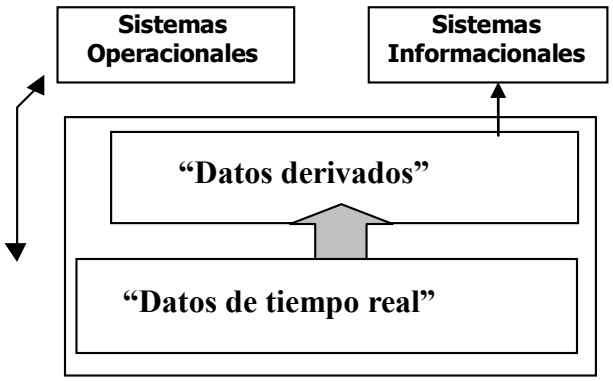


Figura 2.2 Arquitectura de dos capas

La transformación llevada a cabo de los datos de tiempo real a datos derivados requiere de una capa intermedia para solucionar los problemas de inconsistencias, realizando el procesamiento de los distintos conjuntos de datos de tiempo real adecuadamente, a esta capa se le conoce como capa de datos reconciliados y esta nueva arquitectura recibe el nombre de Arquitectura de Datos de **Tres Capas**. Figura2.3

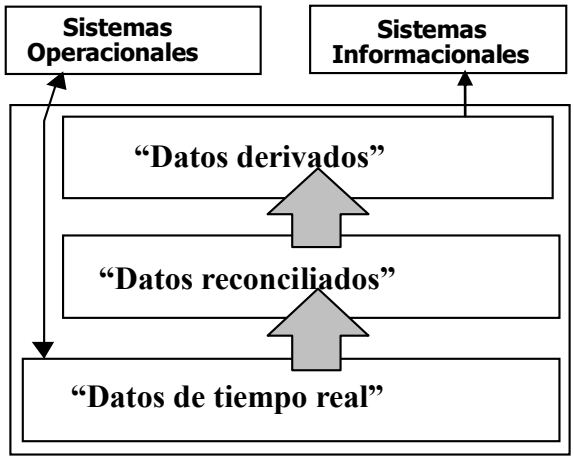


Figura 2.3 Arquitectura de tres capas

Para realizar el diseño de un Data Warehouse se puede seleccionar cualquiera de las arquitecturas anteriormente mencionadas de acuerdo a las características y necesidades de la empresa.

2.4 Arquitectura conceptual de un Data Warehouse

El desarrollo de los data warehouse actualmente llevan un crecimiento acelerado debido a que es una tecnología muy entendible y para comprender cómo se relacionan todos sus componentes es necesario contar con un modelo de Arquitectura Data Warehouse. Considerando su estructura en un marco de ocho niveles:

1. **Nivel operacional:** Se definen los orígenes de datos del almacén, es decir las diversas bases de datos operacionales y fuentes externas de donde serán extraídos.

2. **Nivel de acceso a la información:** Incluye las herramientas de consultas, análisis, generadores de informes y herramientas de data mining, teniendo como finalidad la manipulación, análisis y presentación de los datos de acuerdo a los requerimientos de los usuarios. Cuya finalidad es servir de soporte a las decisiones gerenciales.

3. **Nivel de acceso a los datos:** Es el encargado de la conexión entre el nivel de acceso a la información y el nivel operacional, siendo SQL el lenguaje de datos estándar para el intercambio de los mismos.

4. **Nivel de directorio de datos o Metadata:** Es un repositorio para almacenar y gestionar los metadatos. Este repositorio y su diseño son aspectos de suma importancia para el éxito de un data warehouse, aunque su valor en los proyectos de desarrollo ha sido subestimado (SACHDEVA 1998). Su importancia radica en el hecho de que todo el conocimiento sobre la creación de un data warehouse es almacenado en el mismo y de aquí que los metadatos sean los responsables de guiar los procesos de extracción, limpieza y carga de los datos dentro del almacén además de ayudar a que las herramientas de consulta y los generadores de informe funcionen correctamente (KIMBALL 1998), (MARCO 2000). Los metadatos se refieren a la información sobre la estructura, el contenido y las interdependencias que existen entre los componentes del data warehouse. Estos describen los tipos de datos, las definiciones físicas y lógicas de los mismos, las consultas e

informes predefinidos, las reglas de validación y negocio, las definiciones de las fuentes de datos, las rutinas de transformación y de proceso, etc. En definitiva, se refieren a cualquier cosa que define un objeto del data warehouse.

5. **El Nivel de gestión de proceso:** Es el encargado de la programación de diversas tareas que deben realizarse para construir y mantener el data warehouse y la información del directorio de datos. Es el encargado de controlar varios procesos con el propósito de conservar actualizado el Data warehouse.

6. **El Nivel de mensaje de la aplicación:** Tiene que ver con el envío de información alrededor de la red de la empresa y puede usarse para recolectar las transacciones o los mensajes y entregarlos a una ubicación segura en un tiempo seguro.

7. **Nivel de data warehouse:** Es donde ocurre el almacenamiento físico de datos, usada principalmente para usos estratégicos, de manera que los datos almacenados sean flexibles y fáciles de acceder. En algunos casos puede verse el data warehouse simplemente como una vista lógica, pues puede no involucrar almacenamiento de datos.

8. **Nivel de organización de datos:** Es el componente final de la arquitectura data warehouse, conocido también como gestión de copia o réplica, incluye procesos para combinar, cargar datos para el depósito, resumir, acceder a la información desde bases de datos operacionales y/o externas, permite además el análisis de calidad de datos y filtros que identifican modelos y estructura de datos dentro de la data operacional existente.

2.5 OLTP VS Almacenamiento dimensional de datos

El procesamiento de transacciones en líneas (OLTP) es diferente al almacenamiento dimensional de datos, en cuanto a usuarios, contenido, estructuras de datos, hardware, software, administración, manipulación de los sistemas y ritmos diarios. A pesar de estas diferencias, se continúa usando su pensamiento y herramientas, para diseñar las bases de datos del almacén de datos.

Ralph Kimball (KIMBALL 1996) plantea que las técnicas y los instintos de diseño apropiados para el procesamiento de transacciones son inapropiados y destructivos para el almacenamiento de información, por tal motivo propone un conjunto de diferentes técnicas, denominadas modelación dimensional.

2.5.1 Consistencia.

La consistencia de los datos tanto en los sistemas OLTP como en el almacén de datos dimensional es necesaria, pero es tratada de diferentes formas teniendo en cuenta el propósito de cada sistema.

En OLTP la consistencia de los datos tiene que ser garantizada a nivel de transacción, centrándose en procesar un gran número de transacciones atómicas y pequeñas sin perder nada de ellas. Sin embargo en un almacén de datos dimensional la consistencia se mide de forma global cuidando que la carga de los datos nuevos sea un conjunto completo y consistente, refiriéndose a esto Kimball dice: "…en lugar de una perspectiva microscópica, se tiene una perspectiva de aseguramiento de la calidad. En lugar de un cálculo técnico de la consistencia de los datos, se tiene un juicio directivo de la consistencia de los datos" (KIMBALL 1996).

2.5.2 La Dimensión Tiempo.

En los sistemas OLTP y en los almacenes de datos se trata el tiempo de forma diferente. Las bases de datos OLTP son conocidas como parpadeantes ya que sus datos cambian constantemente, imposibilitando que se pueda representar correctamente la historia anterior, pues al tenerse una gran cantidad de transacciones que alteran la historia, es casi imposible reconstruir rápidamente la instantánea de un negocio en un punto específico en el tiempo.

A diferencia de las bases de datos OLTP, el almacén de datos dimensional no es parpadeante, sólo cambia durante el proceso de carga de datos y es estable cuando los usuarios lo están consultando, garantizando así la consistencia de los datos y si además la información es almacenada cuidadosamente en cada instantánea se puede representar todos los puntos anteriores en el tiempo correctamente.

La instantánea es conocida como el extracto de los datos productivos, migrándose ese extracto al sistema almacén de datos al mismo tiempo cada día, o cada semana, o cada mes como parte de la carga de datos productivos. Este proceso da ascenso a dos fases muy diferentes que todo almacén de datos comparte: la carga y la consulta. (KIMBALL 1996)

2.5.3 El Modelo de Datos Entidad –Relación.

La diferencia más significativa entre ambos sistemas es el modelo de datos, es decir la forma en que estos se organizan. Los sistemas OLTP utilizan el modelo entidad – relación el cual busca eliminar la redundancia de los datos, para posibilitar que las transacciones que alteran el contenido de los mismos sólo toquen a la base de datos en un punto y así mejorar los tiempos de procesamiento. Este modelo presenta una gran simetría, no permitiendo diferenciar qué tabla es más importante o cuál contiene medidas numéricas o descriptores estáticos de los objetos del negocio. Basándose en este modelo se construyen diagramas complejos, debido a que divide los datos en muchas entidades las cuales son convertidas en tablas; las consultas que resumen muchos registros y muchas tablas también se tornan difíciles de comprender por los usuarios y complejas de recorrer por los gestores, por tal motivo Ralph Kimball plantea que: “Los modelos entidad – relación, son un desastre para las consultas debido a que ellos no pueden ser comprendidos por los usuarios y ellos no pueden ser útilmente recorridos por el software del SGBD. Los modelos entidad – relación, no pueden ser usados como bases para el almacén de datos empresarial.”(KIMBALL 1996).

2.5.4 El Modelo Dimensional.

La modelación dimensional es un nuevo nombre para una técnica antigua que permite hacer simples y comprensibles bases de datos, la cual puede ser visualizada como un “cubo” de tres, cuatro, cinco o más dimensiones, donde cualquier punto interior es una intersección de las coordenadas definidas por los ejes del cubo (KIMBALL 1996). Un cubo es la unidad de almacenamiento de información, equivalente a las "tablas" de las bases de datos relacionales, representan la información como matrices.

A los ejes de la matriz se les llama *dimensiones* y representan los criterios de análisis, y a los datos almacenados en la matriz se los llama *medidas* y representan los indicadores o valores a analizar.

La principal característica del modelo dimensional es su sencillez, permitiéndoles a los usuarios una fácil comprensión de las bases de datos, además de posibilitar al software un recorrido eficiente de sus estructuras. Este modelo también es llamado comúnmente entre los diseñadores como **esquema en estrella** debido a su semejanza con una estrella, está compuesto por una tabla principal llamada **hechos** y un conjunto de pequeñas tablas alrededor de esta, nombradas **dimensiones**.

La tabla de hechos está compuesta por múltiples enlaces que la conectan con el resto y por mediciones numéricas del negocio. Cada una de las mediciones es tomada como la intersección de todas las dimensiones; las tablas dimensionales son aquellas donde se almacenan las descripciones textuales de las dimensiones del negocio, cada una de estas descripciones ayuda a describir un miembro de la dimensión respectiva.

2.6 Modelado de datos

Tanto en las bases de datos tradicionales como en los Data Warehouse existen tres niveles de modelado de datos: conceptual, lógico y físico. Las diferencias entre los DWH con las bases de datos operacionales en cuanto al tipo de consultas y rendimiento esperado, hacen que las estrategias de diseño y los modelos de datos utilizados para el DW sean diferentes.

2.6.1 Modelo de datos conceptual

El modelo conceptual captura la información fundamental acerca de las entidades del dominio del problema y sus relaciones. Este modelo es más cercano al espacio del problema que al espacio de la solución.

2.6.2 Modelo de datos lógico

El modelo lógico describe los datos en detalle, generalmente incluye todas las entidades y relaciones entre ellas, sus atributos y tipos de datos, así como las llaves primarias y extranjeras, sin tener en cuenta cómo ellos se implementarán físicamente en la base de datos. Es un puente entre el nivel conceptual y el físico.

2.6.3 Modelo de datos físico

Este modelo describe las estructuras de almacenamiento y los métodos usados para tener un acceso efectivo a los datos.

2.7 Metodologías Utilizadas

En la actualidad existen varias propuestas de modelos y métodos para llevar a cabo el diseño de almacenes de datos, tales como (CABIBBO and TORLONE 1998) , (GOLFARELLI and RIZZI 1998), (BLASCHKA *et al.* 1998), (TRYFONA *et al.* 1999), (TRUJILLO *et al.* 2001), (ABELLÓ *et al.* 2002), sin embargo ninguno de estas propuestas han sido aceptadas como un modelo estándar para realizar el modelado multidimensional, pues no cubren todas las fases y transformaciones necesarias. Debido a esta gran variedad de modelos utilizados en las fases de diseño de los DW se desarrolló un método que proporciona guías de diseño para crear y transformar estos modelos durante la fase de desarrollo del almacén datos, *el Data Warehouse Engineering Process (DWEP)*, propuesto en la tesis de Sergio Luján Mora, (LUJAN-MORA 2005), el mismo es un método orientado a objetos, independiente de cualquier implementación específica, ya sea relacional, multidimensional, orientado a objetos, etc. permite la representación de todas las etapas del diseño de un data warehouse, está basada en UML que es el Lenguaje Unificado de Modelado (OMG 2003) y RUP que es el Proceso Unificado de Desarrollo de Software "…Rup es más que un simple proceso; es un marco de trabajo genérico que puede especializarse para una gran variedad de sistemas software, para diferentes áreas de aplicación, diferentes tipos de organizaciones, diferentes niveles de aptitud y diferentes tamaños de proyectos" (JACOBSON *et al.* 2000).

El método *DWEP* propone la estructuración del almacén de datos en cinco etapas y tres niveles.

Etapas:

- Origen: Define los orígenes de datos del almacén de datos, como los sistemas OLTP, fuentes de datos externas, etc.
- Integración: Define el mapeo entre los orígenes de datos y el propio almacén de datos.
- Almacén de Datos: Define la estructura del almacén de Datos.
- Adaptación: Define el mapeo entre el almacén de datos y las estructuras empleadas por el cliente.
- Cliente: Define las estructuras concretas que son empleadas por los clientes para acceder al almacén de datos, como Data Marts o aplicaciones OLAP.

Niveles:

- Conceptual: Define el almacén de datos desde un punto de vista conceptual, es decir, desde el mayor nivel de abstracción y contiene únicamente los objetos y relaciones mas importantes.
- Lógico: Abarca aspectos lógicos del diseño del almacén de datos, como la definición de las tablas y claves, la definición de los procesos ETL (Extraction*, Transformation and Loading)*, etc.
- Físico: Define los aspectos físicos del almacén de datos, como el almacenamiento de las estructuras lógicas en diferentes discos o la configuración de los servidores de bases de datos que mantienen el almacén de datos.

Como el *DWEP* es una instanciación de RUP para el desarrollo de almacenes de datos establece al igual que este, el ciclo de vida de un proyecto en cuatro fases: Inicio, Elaboración, Construcción y Transición, así como cinco flujos de trabajo fundamentales: Requerimientos, Análisis y Diseño, Implementación y Prueba (JACOBSON *et al.* 2000), además adiciona dos nuevas actividades: Mantenimiento y Revisión Pos-Desarrollo.

En cada flujo de trabajo se utilizan diferentes diagramas UML, para modelar y documentar el proceso de desarrollo.

Para el diseño del GestCon Mart se utilizará dicho método (DWEP), teniendo como principal ventaja el empleo de la misma notación (basada en UML) para el diseño de los diferentes diagramas y las correspondientes transformaciones de una manera integrada. Como UML es un lenguaje de modelado general se utilizan sus mecanismos de extensión para adaptarlo al dominio específico de los almacenes de datos, además se empleará la herramienta Case de Rational (Rational Rose).

2.8 Gestores de bases de datos

Existen diferentes gestores de bases de datos que dan soporte a los sistemas OLTP para soportar las transacciones diarias de las empresas. Estos gestores han venido perfeccionándose desde los años 80, pero no todos han dado soporte a los sistemas OLAP (Procesos Analíticos en Línea) para la exploración de datos y para la extracción de conocimiento. Desde el punto de vista teórico un sistema OLAP debe cumplir las reglas del Dr. Codd (CODD *et al.* 1993):

1. Vista conceptual multidimensional de los datos (facilita el análisis y diseño de modelos de decisión).
2. Transparencia respecto al usuario (la complejidad del sistema no debe ser percibida por el usuario, con el fin de no disminuir su productividad).
3. Facilidad de acceso (el análisis ha de poder realizarse sobre datos provenientes de fuentes de datos heterogéneas, ofreciendo una vista unificada, coherente y consistente de los mismos).
4. Desempeño consistente de reportes (el desempeño no debe degradarse al aumentar la cantidad de dimensiones).
5. Arquitectura cliente / servidor (el sistema OLAP ha de funcionar en un entorno cliente/servidor).
6. Dimensiones simétricas (todas las operaciones han de permitirse sobre cualquiera de las dimensiones).
7. Manejo óptimo de matrices poco densas (el almacenamiento físico de los datos ha de ajustarse a la distribución de sus valores).
8. Soporte multi-usuario (las herramientas OLAP deben soportar el acceso concurrente de varios usuarios).
9. Operaciones cruzadas entre dimensiones sin restricción (las herramientas deben manejar los cálculos y no requerir que el usuario los defina)
10. Manipulación intuitiva de datos (cada dimensión debe contener toda la información que necesite el usuario para efectuar cualquier acción inherente)
11. Reportes flexibles (posibilitar al usuario hacer ajustes)
12. Dimensiones y niveles de adición ilimitados (una herramienta OLAP debería permitir la definición de modelos multidimensionales con 15 dimensiones como mínimo con un número ilimitado de niveles de agregación).

Hoy en día los gestores están actualmente lidereados por dos de las compañías de más experiencias en el soporte a los sistemas OLTP: Microsoft SQL Server y Oracle, que además han sido pioneras en proporcionar herramientas para el diseño de los Data Warehouse.

2.8.1 Microsoft SQL Server 2000

Gracias a su escalabilidad Microsoft SQL Server 2000 puede reducir drásticamente el tiempo necesario para poner en funcionamiento aplicaciones de data warehousing y aplicaciones de negocio, utilizando las herramientas interactivas proporcionadas por las versiones de SQL Server Personal, Standard y Enterprise.

Es una plataforma para bases de datos de proceso transaccional en línea (OLTP), almacenamiento de datos y aplicaciones de comercio electrónico, facilita Servicios de transformación de datos (DTS, *Data Transformation Services*) utilizándose para administrar el flujo de datos entrante en la base de datos relacional, así como el flujo que sale de esta hacia los cubos de Analysis Services de SQL Server 2000, para la creación de informes y consultas interactivas. Analysis Services es una herramienta brindada por este gestor, el cual es un servidor de nivel intermedio para procesos analíticos en línea y minería de datos, también incluye un servidor para la administración y análisis de los cubos de datos multidimensionales, proporcionando un rápido acceso a la información almacenada.

Analysis Services proporciona una arquitectura cliente servidor que facilita un rápido acceso a los datos y una arquitectura para obtener acceso a los datos de la minería de datos. Organiza los datos en cubos, creando datos de agregación precalculados para proporcionar respuestas rápidas a consultas analíticas complejas, permite la creación de modelos de minería de datos de orígenes multidimensionales (OLAP) y relacionales, proporciona un servicio llamado PivotTable (DALGLEISH 2006) para que aplicaciones como Microsoft Excel y otras aplicaciones de diferentes fabricantes puedan recuperar datos del servidor para ser mostrados al usuario o para la creación de cubos de datos locales, es compatible con varios modelos de almacenamiento como: MOLAP, ROLAP y HOLAP.

2.8.1.1 Modelos de almacenamiento de datos

Las opciones de almacenamiento físico afectan al rendimiento, a los requisitos de almacenamiento y a las ubicaciones de almacenamiento de las particiones y de sus cubos primarios. Una de estas opciones es el

modo de almacenamiento de la partición. Una partición puede tener uno de estos tres modos de almacenamiento básicos:

- OLAP multidimensional (MOLAP): El modo de almacenamiento MOLAP da lugar a que las agregaciones de la partición y una copia de sus datos de origen se almacenen en una estructura multidimensional de alto rendimiento, por tanto las consultas pueden resolverse sin obtener acceso a los datos de origen de la partición, incluso cuando no se puedan obtener los resultados de las agregaciones de la partición. El modo de almacenamiento MOLAP proporciona los tiempos de respuesta de consulta más rápidos posibles, en función del diseño y porcentaje de las agregaciones de la partición, así como un rendimiento y una compresión de los datos excelente. En general, el modo MOLAP es más apropiado para las particiones de cubos que se utilizan con frecuencia y donde se necesita una rápida respuesta de consultas.

- OLAP relacional (ROLAP): Este modo de almacenamiento hace que las agregaciones de la partición se almacenen en tablas de la base de datos relacional especificada en el origen de datos de la partición, permitiendo sacarle las ventajas a la tecnología relacional. A diferencia del modo de almacenamiento MOLAP, no da lugar a que se almacene una copia de los datos de origen; cuando los resultados no pueden derivarse de las agregaciones o de la caché de cliente, se obtiene acceso a la tabla de hechos de la partición para responder a las consultas. Con este modo de almacenamiento, la respuesta de consultas suele ser más lenta que la de los otros dos modos. ROLAP se usa normalmente para el acceso a grandes conjuntos de datos que se consultan con poca frecuencia, tales como datos históricos de años no recientes.

- OLAP híbrido (HOLAP): HOLAP combina atributos de los modos MOLAP y ROLAP. Al igual que MOLAP, almacena los datos de las agregaciones de la partición en una estructura multidimensional de alto rendimiento, pero no almacena una copia de los datos de origen. HOLAP es el equivalente de MOLAP para las consultas que sólo obtienen acceso a los datos de resumen contenidos en las agregaciones de una partición. Las consultas que obtienen acceso a los datos de origen deben recuperar datos de la base de datos relacional y no serán tan rápidas como lo serían si los datos de origen se hubieran almacenado en la estructura MOLAP. Las particiones almacenadas como HOLAP son más pequeñas que sus equivalentes MOLAP y responden más rápidamente que las particiones ROLAP a las consultas que implican datos de resumen. El modo

de almacenamiento HOLAP suele ser más adecuado para particiones en cubos que requieren una respuesta de consultas rápida basada en una gran cantidad de datos de origen (MICROSOFT 2000).

El modo de almacenamiento que se utilizará en el GestCon Mart todavía no se ha precisado, el mismo se definirá más adelante cuando el Data Mart esté funcionando y se pueda determinar cuáles de los cubos son los más visitados.

2.8.1.2 Requisitos mínimos de hardware y software

- El procesador debe ser Pentium o superior.
- Velocidad de procesador: 166 MHz.
- Memoria física RAM: 64 MB, aunque el Personal puede tener 32 MB en los sistemas operativos que no sean de la familia Windows 2000.
- Espacio en disco duro:
  - Componentes de base de datos de SQL Server: De 95 a 270 MB, 250 MB típica.
  - Analysis Services: 50 MB mínimo, 130 MB típica.
  - English Query: 80 MB.
  - Desktop Engine: 44 MB.
- Sistemas Operativos que lo soportan: Todas las ediciones de Microsof SQL Server 2000 se pueden ejecutar a partir del sistema operativo Windows NT Server, a excepción de que en la familia de Windows 2000 en la versión Professional Edition se ejecutan Personal y Developer solamente**,** este último también es soportado en Windows NT Workstation y el Personal además puede correr en Windows 98 y Windows Me. Si solo se desea instalar las herramientas cliente se pueden utilizar Windows NT 4.0, Windows Me, Windows 98 y Windows 2000 (todas las versiones) o superiores (MICROSOFT 2000).

2.8.2 Microsoft SQL Server 2005

SQL Server 2005 está construido sobre las fortalezas de SQL Server 2000, aumenta el rendimiento, la confiabilidad, la disponibilidad, la capacidad de programación y la facilidad de uso del SQL Server 2000, ofrece nuevas oportunidades de diseño para los desarrolladores de cubos y proporciona un enfoque nuevo para el uso de los sistemas OLAP.

Brinda la herramienta Microsoft SQL Server 2005 Analysis Services (SSAS) a la cual se le han agregado nuevas características, además de las ya existentes en Análisis Services. Esta herramienta incluye mejoras como métricas empresariales personalizables en los cubos, denominadas indicadores clave de rendimiento (KPI); la creación de varias tablas de hechos en un único cubo; medidas de suma parcial para agregar medidas a una dimensión y no a otras; mejoras en las dimensiones como atributos, pues en versiones anteriores las dimensiones se basaban directamente en los niveles de una jerarquía y ahora se basan en atributos que corresponden a las columnas de las tablas de dimensión, separando así las características estructurales de una dimensión de su característica de exploración; relaciones de dimensiones de referencia, admitiéndose las dimensiones de referencia mediante el uso de relaciones entre las dimensiones de referencia y un grupo de medida de otra dimensión lo cual posibilita asociar estas dimensiones a un cubo sin crear un esquema de copo de nieve; además brinda un tamaño de dimensiones prácticamente ilimitado puesto que ya no se depende del almacenamiento residente en memoria.(TANG and MACLENNAN 2005)

Microsoft SQL Server 2005 brinda una nueva plataforma para crear soluciones de integración de datos de alto rendimiento que se llama Microsoft SQL Server 2005 Integration Services (SSIS) que sustituye a los Servicios de transformación de datos (DTS) del SQL Server 2000. Incluye paquetes de extracción, transformación y carga (ETL) para el almacenamiento de datos; resuelve muchas dificultades y limitaciones de los DTS incluyendo mejoras como una nueva arquitectura extensible, un nuevo diseñador de paquetes, estructuras de bucle y transformaciones, así como mejoras en la implementación, administración y el rendimiento de los paquetes; nuevas herramientas gráficas y asistentes para crear y depurar paquetes; tareas para realizar funciones de flujo de trabajo como la ejecución de comandos SQL, operaciones FTP y mensajería por correo electrónico; transformaciones para borrar, agregar, mezclar y copiar datos; servicio de administración e interfaces de programación de aplicaciones (API) para programar el modelo de objetos de Integration Services.

Entre las nuevas tareas proporcionada por Integration Services se encuentran las de flujo de trabajo como la ejecución de otros paquetes y aplicaciones, el envío de mensajes de correo electrónico; tarea de preparación de datos como la carga, descarga y copia de archivos; tarea sistema de archivos, para realizar operaciones en archivos y carpetas del sistema de archivos; tarea Servicio Web y XML; tarea de nuevos orígenes de datos y destinos, pues además de los de SQL Server, OLE DB y de los archivos planos se brinda destino de SQL Server Mobile para insertar y actualizar datos en las bases de datos de SQL Server Mobile, origen y destino de DataReader para consumir y proporcionar datos a cualquier proveedor de datos .NET Framework, origen XML, origen y destino de archivo sin procesar. (MICROSOFT 2007)

Los requerimientos mínimos de hardware y software para las diferentes versiones del SQL Server 2005: Enterprise Edition, Developer Edition, Standard Edition, Workgroup Edition, Express Edition, son mucho más elevados que para las versiones de SQL Server 2000.

2.8.2.1 Requisitos mínimos de hardware y software

- Se requiere un procesador compatible con Pentium III o superior para todas las versiones.
- Velocidad de procesador: 600 MHz aunque se recomienda 1GHz o superior, para todas las versiones.
- Memoria física (RAM): Las versiones Enterprise, Developer, Standard y Workgroup 512 MB, aunque se recomienda 1 GB o superior y el Express: 192 MB pero es recomendado 512 MB o más.
- Espacio en disco duro: Disponer de 1,6 GB de espacio en la unidad del sistema. Este requisito es aplicable incluso si se instala componentes de SQL Server en una unidad distinta de la predeterminada.
- Sistemas operativos que ejecutan el software de servidor: Todas las versiones de SQL Server 2005, excepto Enterprise, se pueden ejecutar en todos los Sistemas Operativos, menos Windows XP Home Edition y Windows 2003 Web Edition, que solo soportan a Express y a Developer. Como caso particular se tiene a Enterprise que sólo puede ser ejecutado en los siguientes sistemas operativos: Windows 2000 Server, Windows 2000

Advanced Server, Windows 2000 Datacenter Edition, Windows 2003 Server, Windows 2003 Enterprise Edition, Windows 2003 Datacenter Edition, Windows Small Business Server 2003 Standard Edition y Windows Small Business Server 2003 Premium Edition.

2.8.3 Oracle

Ofrece una administración eficiente, confiable y segura de los datos para todo tipo de aplicaciones, desde sistemas de alto volumen de transacciones en línea hasta aplicaciones de consulta intensiva. No sólo soporta las necesidades de la compleja administración de datos, sino que también brinda las herramientas para administrar los sistemas, ofrece la flexibilidad para distribuir efectiva y eficientemente los datos a los usuarios y la escalabilidad para alcanzar el rendimiento óptimo de todos los recursos de computación disponibles, convirtiéndose en el producto líder de data warehousing en el mercado.

Oracle provee una potente herramienta para el diseño de Data Warehouse llamada Oracle Warehouse Builder (OWB). Permitiendo todo el diseño del proceso de extracción, transformación y carga de los datos (ETL) dando soporte a todo el flujo de datos necesarios para poblar los warehouse.

Esta herramienta permite que con una única interfaz fácil de usar se tomen los datos en bruto en diferentes formatos y de diferentes sistemas y sean transformados en información de alta calidad para facilitar la optimización de los reportes y el análisis del negocio. Incorporándole nuevas características como el editor de mapas que facilita numerosas características en la interfaz de usuario y permite un diseño mucho más cómodo, aumentando la productividad y reduciendo el número de los errores. A la última versión se le han incorporado más fuentes de datos, adicionando un número de paquetes de aplicaciones conectores a diferentes productos e incrementando las funcionalidades del ya presente conector SAP, contrario a las demás herramientas ETL que usan SQL y PL/SQL, SAP usa ABAP que es el lenguaje nativo SAP y que corre directamente en el SAP Server lo que posibilita que estos conectores tengan un mejor rendimiento y un incremento en la productividad. Permite crear el diseño lógico describiendo los cubos OLAP en dimensiones, jerarquías, medidas, medidas precalculadas y cualquier otro componente que se necesite, usando el nuevo XML API con la opción OLAP se puede crear un espacio de trabajo analítico y los metadatos requeridos en el catálogo de la base de datos.

Warehouse Builder posibilita que se elija el tipo de implementación, MOLAP (multidimensional OLAP) o en ROLAP (relacional OLAP). (ORACLE 2007)

2.8.3.1 Requerimientos mínimos de hardware y software

- Memoria física (RAM): 256 MB, aunque es recomendado 512 MB.
- Memoria Virtual: El doble de la RAM.
- Velocidad de procesador: 550 MHz.
- Espacio en disco:
  - Instalación Básica: 2.04 GB.
  - Instalación Avanzada: 1.94 GB
- Sistemas operativos que lo ejecutan: Windows 2000 con service pack 1 o superiores, además de todas las versiones de Windows Server 2003 y Windows XP Profesional.(ORACLE 2007)

2.8.4 ¿Por qué Microsoft SQL Server 2000?

Teniendo en cuenta las características de los Gestores de Bases de Datos anteriormente expuestos, se utilizará Microsoft SQL Server 2000 a pesar de las ventajas y de las nuevas tareas incorporadas en el SQL 2005 y de la excelente herramienta proporcionada por Oracle (OWB) para el fácil diseño y la probada eficiencia en el manejo de grandes volúmenes de datos, pues Microsoft SQL Server 2000 ha demostrado ser un gestor estable al cual se le han corregido sus errores, el mismo es capaz de manejar eficientemente todo el volumen de información de nuestro Data Mart y brindarnos todo el soporte necesario para su construcción y mantenimiento sin necesitar grandes recursos de hardware, ni un potente sistema operativo.

2.9 Conclusiones

En este capítulo se realizó un estudio teórico y conceptual sobre Data Warehouse, a partir de un conjunto de libros y artículos especializados. Basándonos en este estudio detallado usaremos para la implementación del GestCon Mart el enfoque de Raplh Kimball, pues es menos costoso, es más funcional, se pueden priorizar las áreas más críticas, además su estructura de datos ofrece una mayor facilidad al usuario para la exploración y búsqueda de información. Consideramos que la mejor y más completa arquitectura de datos es la de tres capas, siempre y cuando la capacidad de almacenamiento no sea un

obstáculo, puesto que resuelve mejor los problemas relacionados con la conciliación de los datos, eliminando las inconsistencias de los mismos y evitando que los usuarios obtengan información errónea y como gestor de Base de Datos se seleccionó Microsoft SQL Server 2000.

3 Capitulo 3: Implementación del GestCon Mart.

3.1 Introducción

En este capítulo se realizará una descripción detallada del GestCon Mart, se tratará el proceso de diseño de los DWH así como la realización de las diferentes fases de diseño del Data Mart utilizando la metodología DWEP a través de varios diagramas creados en tres niveles: conceptual, lógico y físico.

3.2 Descripción detallada del Gestcon Mart

Se desea construir un Data Mart para posibilitar el análisis de los problemas y las soluciones ocurrido en los proyectos productivos de la universidad. El mismo almacenará el nombre del proyecto, la fecha de creación y de cierre, la plataforma que utiliza, el organismo al que pertenece, las tareas orientadas en el proyecto, los problemas presentados en la resolución de las mismas, las soluciones obtenidas, así como otros datos de interés. Todo esto debidamente acotado en el tiempo para garantizar la rápida representación de la historia anterior.

GestCon Mart brindará tres enfoques para la exploración de los datos: los usuarios finales podrán explorar centrándose en el *historial de las personas*, este punto de vista brinda las personas con sus problemas y soluciones. También se podrá buscar los *problemas con sus respectivas soluciones* y la última alternativa se realizará a través de los *problemas y las soluciones con niveles de calidad asociados*, el cual es una evaluación dada tanto a los problemas como a las soluciones. Para dicha exploración se utilizará Microsoft Excel, pues dentro de su funcionalidad nativa posee las tablas dinámicas o pivot tables, las cuales son componentes interactivos que permiten analizar en forma dinámica y muy flexible datos multidimensionales. Ejemplo de su utilización se puede ver en el anexo 3.50, donde se muestra una tabla como resultado de una exploración realizada en términos de personas y problemas, asimismo en el anexo 3.51 se muestra la misma solución, pero esta vez expresado en forma de gráfico, para una mejor representación visual del contenido. En el anexo 3.52 se puede consultar los problemas con sus respectivas soluciones, así como su gráfico en forma piramidal en el anexo 3.52.

3.3 Proceso de diseño de un Data Warehouse

De igual forma que en los sistemas de bases de datos operacionales, el proceso de diseño del DWH puede dividirse en tres etapas secuenciales: diseño conceptual, diseño lógico y diseño físico, (BATINI *et al.* 1992). Pero Las características de los DWH hacen que las estrategias de diseño para las bases de datos operacionales generalmente no sean aplicables para el diseño de DWH (KIMBALL 1996),(INMON, WILLIAM 1996).

En la etapa de diseño conceptual se construye un esquema conceptual de la realidad a partir de los requerimientos y/o bases fuentes, a partir de él se genera un esquema lógico, que es dependiente del tipo de modelo y tecnología de DBMS (Database Management System). Por último, en la etapa de diseño físico se implementa el esquema lógico en el manejador de bases de datos elegido, teniendo en cuenta técnicas de optimización física, como son: índices particiones, etc.

3.4 Aplicación del método DWEP

El DWEP es un método global para llevar a cabo el diseño de todas las fases de los almacenes de datos, incluyendo las fuentes de datos operacionales, los procesos ETL y el propio esquema del almacén de datos, para ello se mostrarán a continuación las fases del ciclo de vida del GestCon Mart con sus respectivos diagramas.

Para la realización de dichos diagramas se hace uso de un *Perfil UML para Modelación Multidimensional* que permite modelar las principales propiedades multidimensionales de un almacén de datos. Este perfil es creado por Sergio Lujan Mora y Juan Trujillo, (TRUJILLO *et al.* 2002), el cual permite llevar a cabo el modelado del DWH utilizando la metodología DWEP.

3.4.1 Requerimientos

“La captura de requisitos es el proceso de averiguar, normalmente en circunstancias difíciles, lo que se debe construir” (JACOBSON *et al.* 2000).

Durante este flujo de trabajo se define el alcance del almacén de datos mediante entrevistas con los usuarios finales. Los requerimientos son modelados usando casos de usos. Una vez definidos los requerimientos, el proyecto DWH es establecido y son designado los diferentes roles.

3.4.1.1 Requerimientos Funcionales

1. El sistema permitirá a los usuarios la presentación de la información por los criterios pre-establecidos.
2. El sistema permitirá alternar las filas y las columnas, correspondientes a las dimensiones en la matriz de datos.
3. El sistema permitirá activar o desactivar dimensiones en las consultas.
4. El sistema tendrá que extraer, transformar y cargar datos de los sistemas operacionales.

3.4.1.2 Requerimientos no Funcionales

1. Rendimiento.

1.1 El tiempo medio de respuesta a consultas simples echas al Data Mart deberá ser de no más de 1 minuto y para las consultas complejas (que implican más de 30 millones de registros), podría ser de 1 a 3 minutos.

1.2 El sistema deberá soportar hasta 500 usuarios activos.

1.3 El sistema deberá soportar hasta 200 accesos a la misma vez.

2. Seguridad.

2.1 El sistema está definido para operar conjuntamente con el sistema del control del acceso y garantizar de esta forma el acceso a los datos solamente a las personas debidamente autorizadas.

3. Accesibilidad.

3.1 El Data Mart debe estar disponible para todos los usuarios las 24 horas del día durante los 7 días de la semana.

4. Precisión de los datos.

4.1 El sistema tendrá que proceder a la actualización periódica de los datos de los proveedores de las fuentes, dentro de los períodos establecidos entre el cliente y equipo de desarrollo.

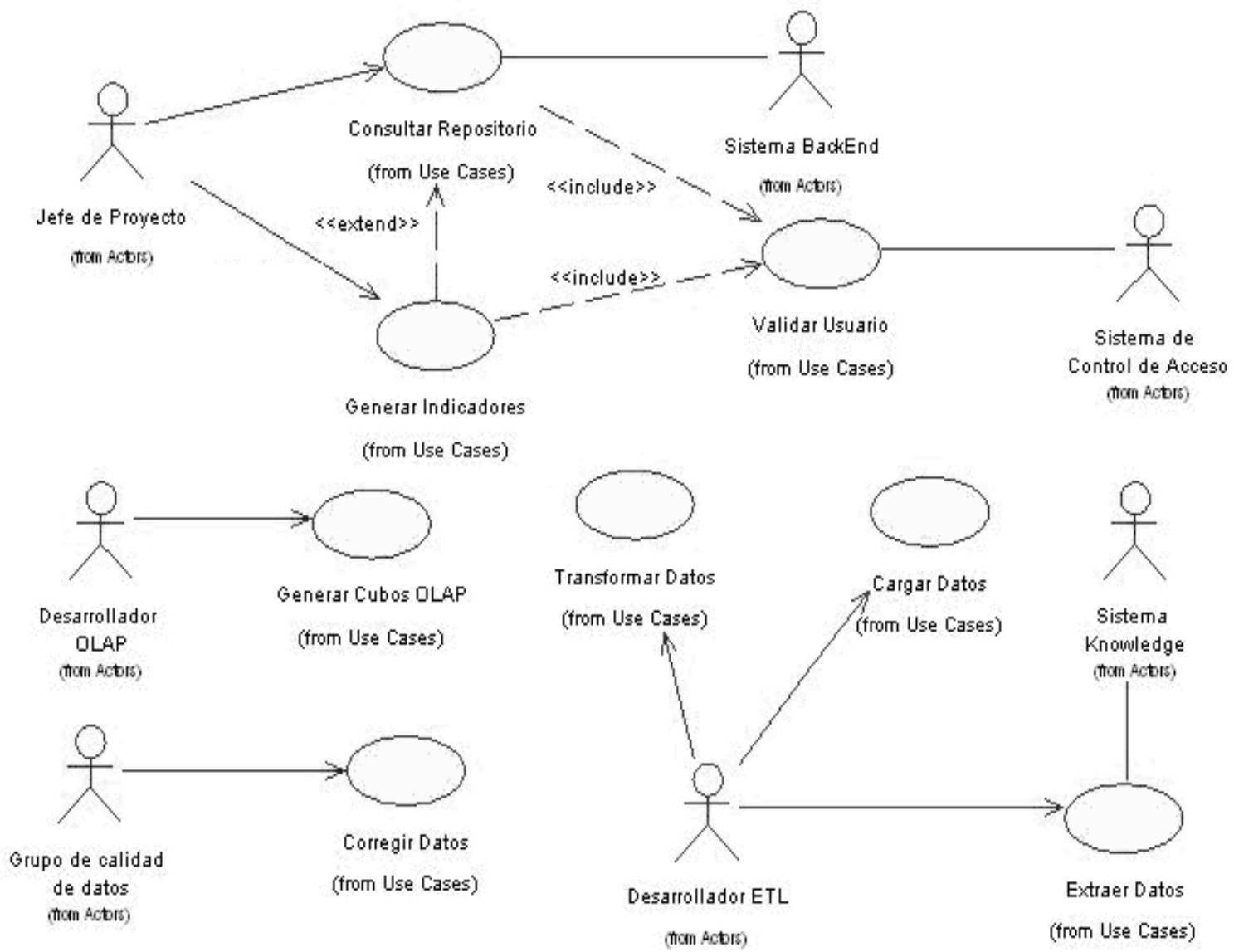
5. Usabilidad.

5.1 El Data Mart deberá ser sencillo, flexible y de fácil uso.

3.4.1.3 Actores del Sistema

Actores del sistema	Justificación
Jefe de Proyecto	Encargado de consultar los datos del Data Mart para llevar a cabo la toma de decisiones en los diferentes proyectos productivos.
Desarrollador ETL	Tiene la tarea de extraer los datos de los sistemas OLTP, transformarlos y cargarlos en el Data Mart.
Desarrollador OLAP	Es el responsable de desarrollar los cubos OLAP.
Grupo de calidad de datos	Este rol tiene la misión de asegurar la exactitud de los datos, llevando a cabo la corrección de los mismos.
Sistema de Control de Acceso	Sistema encargado de controlar el acceso de los usuarios al Data Mart.
Sistema Knowledge	Sistema externo que contiene la base de datos operacional Knowledge, la cual provee los datos al Data Mart.
Sistema BackEnd	Sistema externo encargado de permitir la exploración de datos del GestCon Mart a los usuarios finales. (En este caso se usará Microsoft Excel).

3.4.1.4 Diagramas de casos de uso del sistema



3.4.2 Análisis

El objetivo de este flujo es refinar y estructurar los requerimientos obtenidos en el flujo de trabajo anterior, así como definir las fuentes de datos operacionales y externas que alimentarán el almacén de datos, teniendo en cuenta las necesidades expresadas por los usuarios finales. En la modelación de las fuentes

de datos en diferentes niveles de detalles se realizará el **esquema conceptual (SCS)**, y **el esquema lógico (SLS).**

Para realizar la población de datos hacia el GestCon Mart se parte de una Base de Datos Operacional: Knowledge, la cual se encuentra normalizada y almacena datos sobre los proyectos productivos, su horario de trabajo, la plataforma utilizada, los lenguajes de programación, las personas que conforman el equipo de trabajo, su nivel profesional, el rol desempeñado por cada miembro, las tareas asignadas, los problemas presentados en la resolución de dichas tareas así como las soluciones encontradas a un determinado problema, también se guardarán los datos pertenecientes a la persona encargada de darle una evaluación a las soluciones y problemas propuestos.

3.4.2.1 Esquema conceptual de la fuente (SCS)

El esquema conceptual de la fuente representa a la base de datos operacional que proveerá datos al DWH y tiene como objetivo conocer qué datos están disponibles para el DM, para ello se representan las entidades más importantes y las relaciones entre ellas, permitiendo entender el dominio del mundo real del problema. Ver figura 3.1

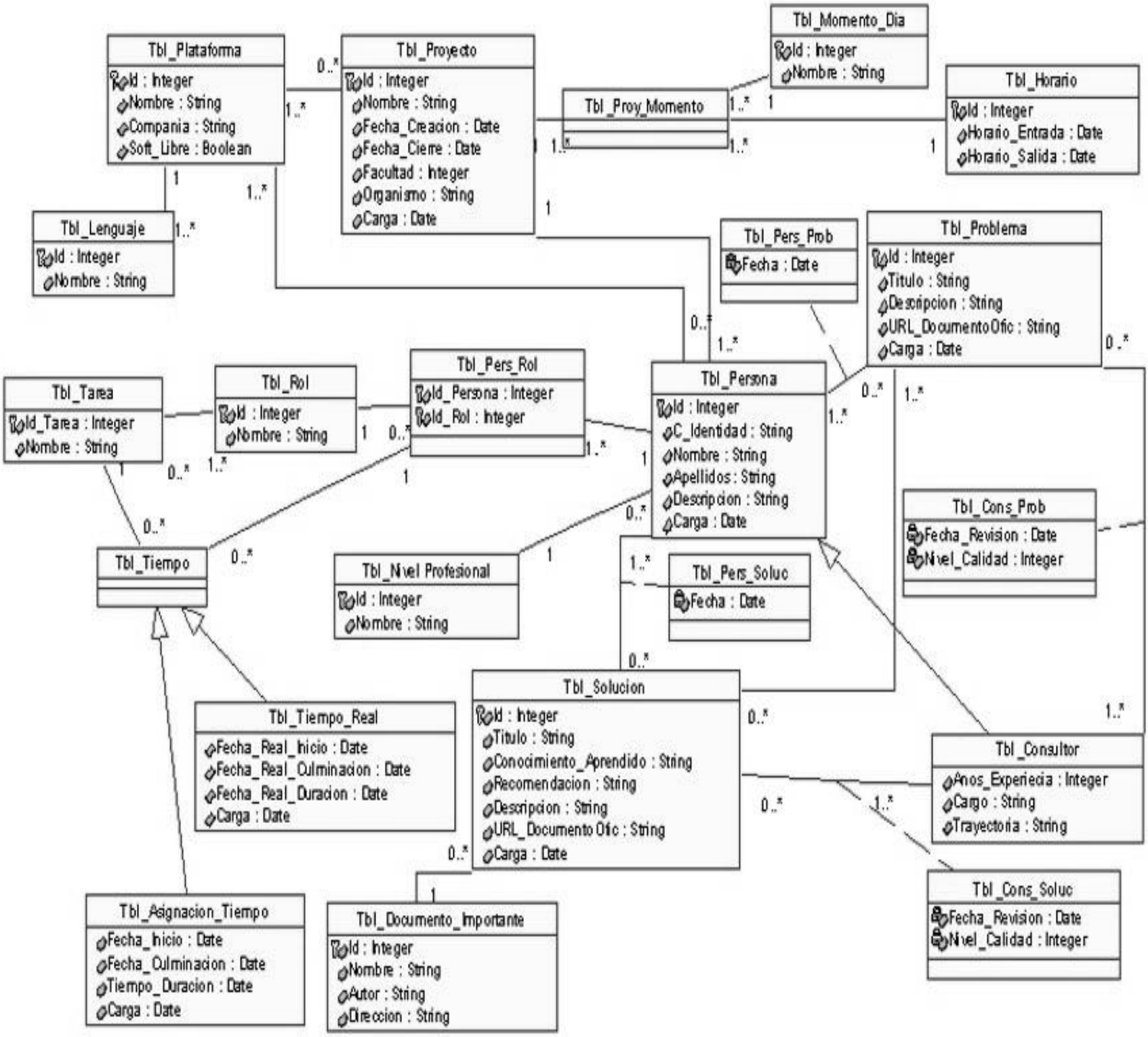


Figura 3.2 Esquema Lógico de la fuente (SCS)

3.4.2.2 Esquema lógico de la fuente (SLS)

Para realizar el esquema lógico de la fuente se toma como entrada el esquema conceptual de la fuente visto anteriormente, para ello las clases son convertidas a tablas, los atributos a columnas y las asociaciones a relaciones. Ver figura 3.2

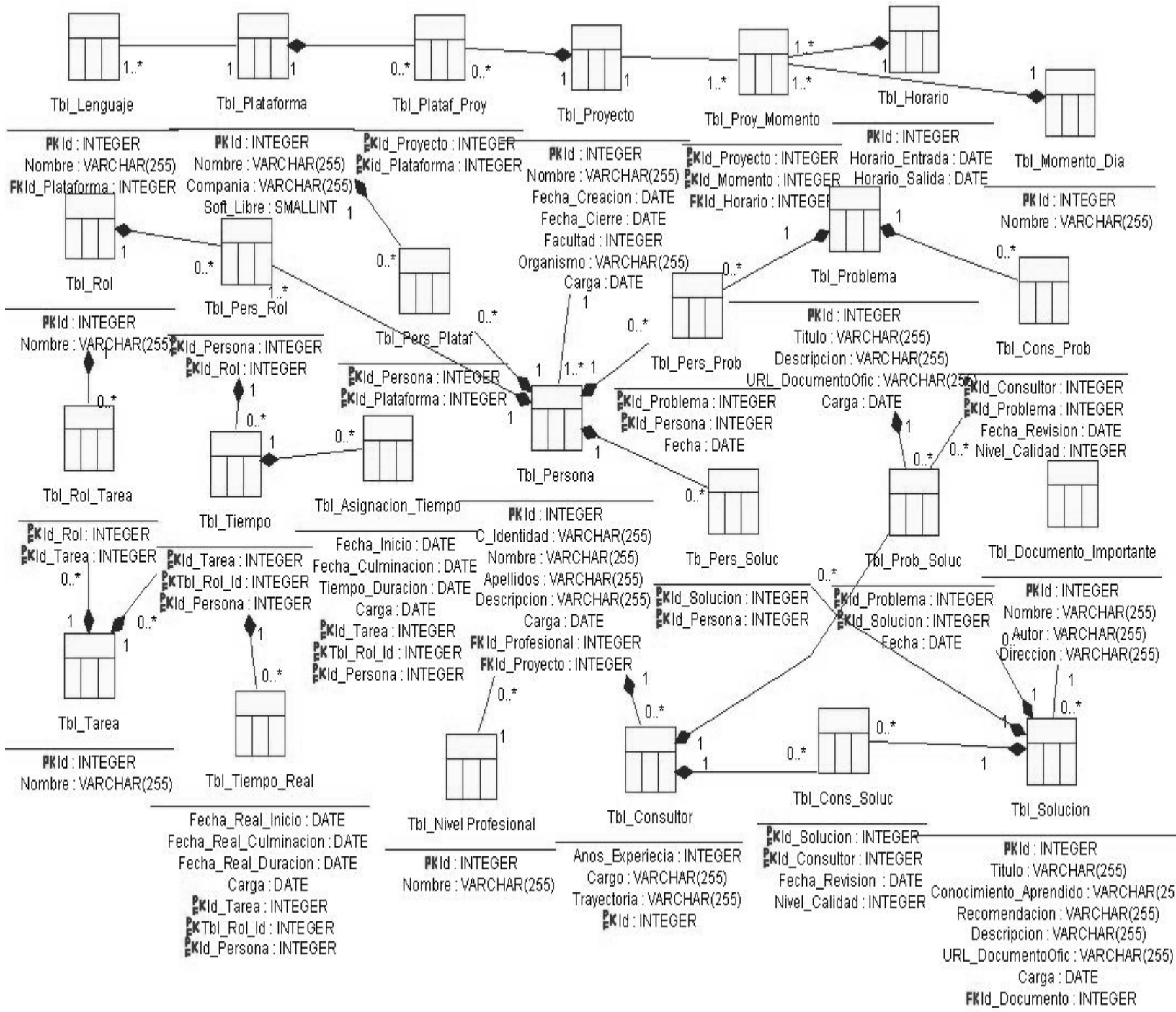


Figura 3.2 Esquema Lógico de la fuente (SLS)

3.4.3 Diseño

En este flujo de trabajo se construye el ***esquema conceptual del DWH (DWCS)***, para realizar dicha actividad existen dos estrategias: *top-down* que recomienda la construcción del DWH primero y luego la de los DM, a diferencia de la propuesta *bottom-up* que usa una serie de DM incrementales para finalmente integrarlos y construir el DWH (HACKNEY 1998),(ECKERSON 2002). En la creación del GestCon Mart se utilizó la estrategia *bottom-up*, porque tiene menos riegos de fracasar, los resultados se ven mucho más pronto y es más probable terminarlo en tiempo.

También se determinan los procesos ETL, los cuales se definen como un mapeo entre las fuentes de datos (esquema conceptual de las fuentes) y el almacén de datos (esquema conceptual del DWH) quedando definido así el ***esquema de mapeo de datos (Data Mapping).***

Como se ha expuesto en las conclusiones del Capítulo1 se usará para la implementación del GestCon Mart el enfoque de Raplh Kimball, quien propone como modelo de datos al dimensional, que estructura la información en hechos y dimensiones.

3.4.3.1 Esquema conceptual del DWH

El diseño conceptual tiene por objetivo la construcción de una descripción abstracta y completa del problema, se seleccionan los objetos relevantes para la toma de decisiones y se especifica el propósito de utilizarlos como dimensiones y/o medidas.

La metodología que se está siguiendo divide el proceso de diseño conceptual del DWH en tres niveles (LUJAN-MORA 2005) para una mejor comprensión:

- Nivel 1: Definición del Modelo: Un Paquete representa un esquema estrella de un modelo multidimensional. En este nivel, una dependencia entre dos paquetes indica que los esquemas estrellas comparten al menos una dimensión.
- Nivel 2: Definición de un esquema estrella. Un paquete representa un hecho o una dimensión de un esquema estrella. En este nivel, una dependencia entre dos paquetes de dimensión indica que las dimensiones comparten al menos un nivel en sus correspondientes jerarquías.

- Nivel 3: Definición de un hecho o una dimensión. Se compone de un conjunto de clases que representan los niveles jerárquicos en un paquete de dimensión o el esquema estrella completo en el caso de un paquete de hecho.

El GestCon Mart se encuentra compuesto por tres cubos de datos: el primero representando al enfoque de exploración de datos *problemas con sus respectivas soluciones,* formado por la tabla de hecho Tbl_Probl_Soluc y a su alrededor las tablas dimensiones Dim_Persona, Dim_Problema, Dim_Proyecto, Dim_Solucion, Dim_Tarea, Dim_Tiempo. El segundo cubo representa el enfoque de *problemas y las soluciones con niveles de calidad asociados,* constituido por la tabla hecho Tbl_Calif_Probl_Soluc con sus pertinentes dimensiones Dim_Persona, Dim_Probl_Sin_CGeneralizada, Dim_Proyecto, Dim_Soluc_Sin_CGeneralizada, Dim_Tarea, Dim_Tiempo. Por último el cubo que da respuesta al enfoque *historial de las personas,* con su tabla hecho Tbl_Historial_Persona y sus dimensiones Dim_Persona, Dim_Problema, Dim_Proyecto, Dim_Solucion, Dim_Tarea, Dim_Tiempo.

En la figura 3.3 se muestra el nivel 1, representando los diferentes cubos de datos que forman al DM, las flechas entre los paquetes denota que tienen dimensiones comunes.

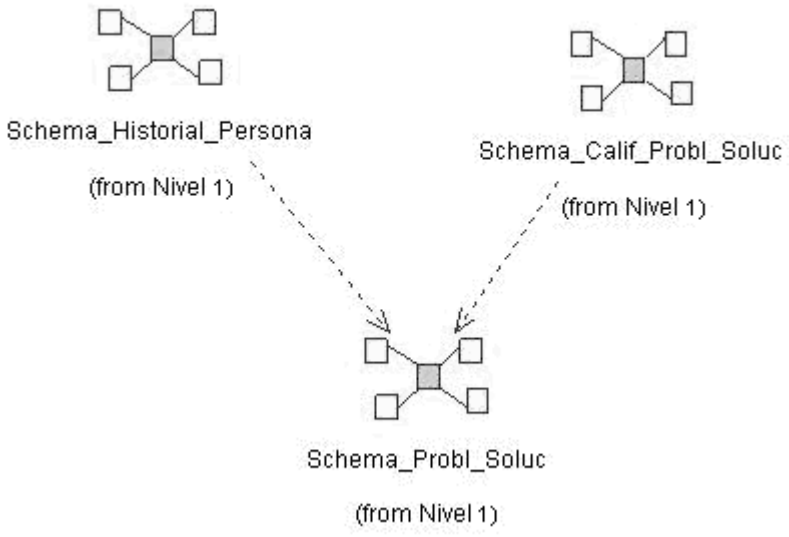


Figura 3.3 Esquema conceptual del DWH. Nivel1

La figura 3.4 muestra el contenido del paquete Schema_Probl_Soluc (nivel 2), donde se observa en el centro un paquete hecho (Pack Fact_Probl_Soluc) que simboliza una tabla hecho y a su alrededor paquetes dimensiones que representan las dimensiones con las cuales está relacionada. El contenido del paquete hecho y de los paquetes dimensiones es representado en el nivel 3. Los diagramas en este nivel están solamente compuestos por clases y relaciones entre ellas.

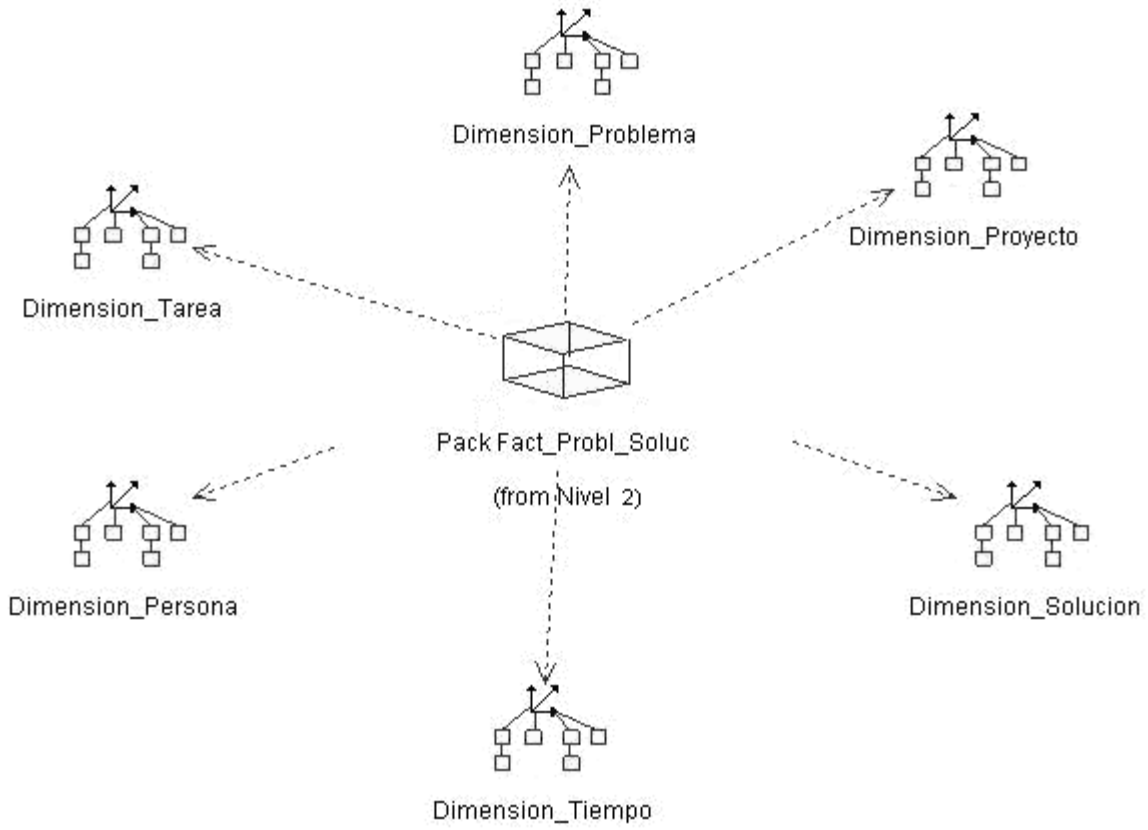


Figura 3.4 Esquema conceptual del DWH. Schema_Probl_Soluc (nivel 2)

En la figura 3.5 se muestra el contenido del paquete Dimension_Persona (nivel 3) que contiene la definición de la clase <<Dimensión>> Persona y los diferentes niveles de jerarquías que son representadas por clases <<Base>>, estos niveles de jerarquías definen como las diferentes operaciones OLAP (roll-up, drill-down) pueden ser aplicadas.

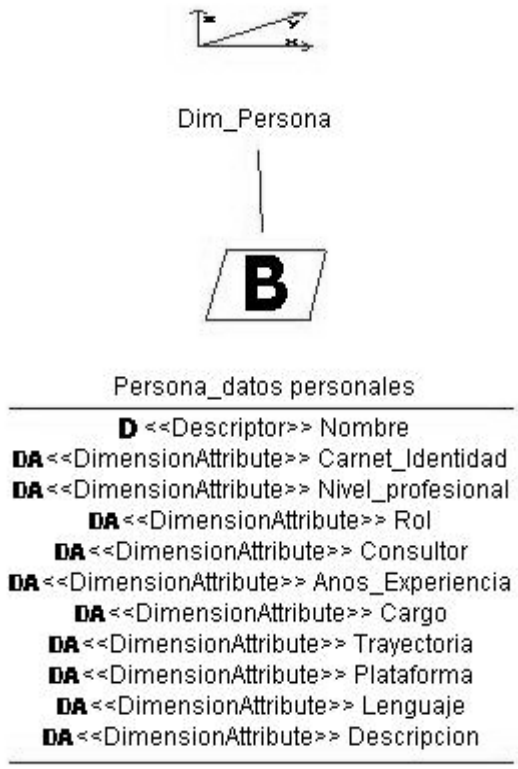


Figura 3.5 Esquema conceptual del DWH. Dimension_Persona (nivel 3)

En los anexos 3.1, 3.2, 3.2, 3.4, 3.5 se muestran los contenidos de las restantes dimensiones: Dimension_Tarea, Dimension_Problema, Dimension_Solucion, Dimension_Proyecto, Dimension_Tiempo, respectivamente.

También se expone en la figura 3.6 el contenido del paquete hecho Pack Fact_Probl_Soluc (nivel 3), en el cual la clase hecho es definida con su correspondiente medida (existencia) y se representa además las clases dimensiones con sus niveles de jerarquías.

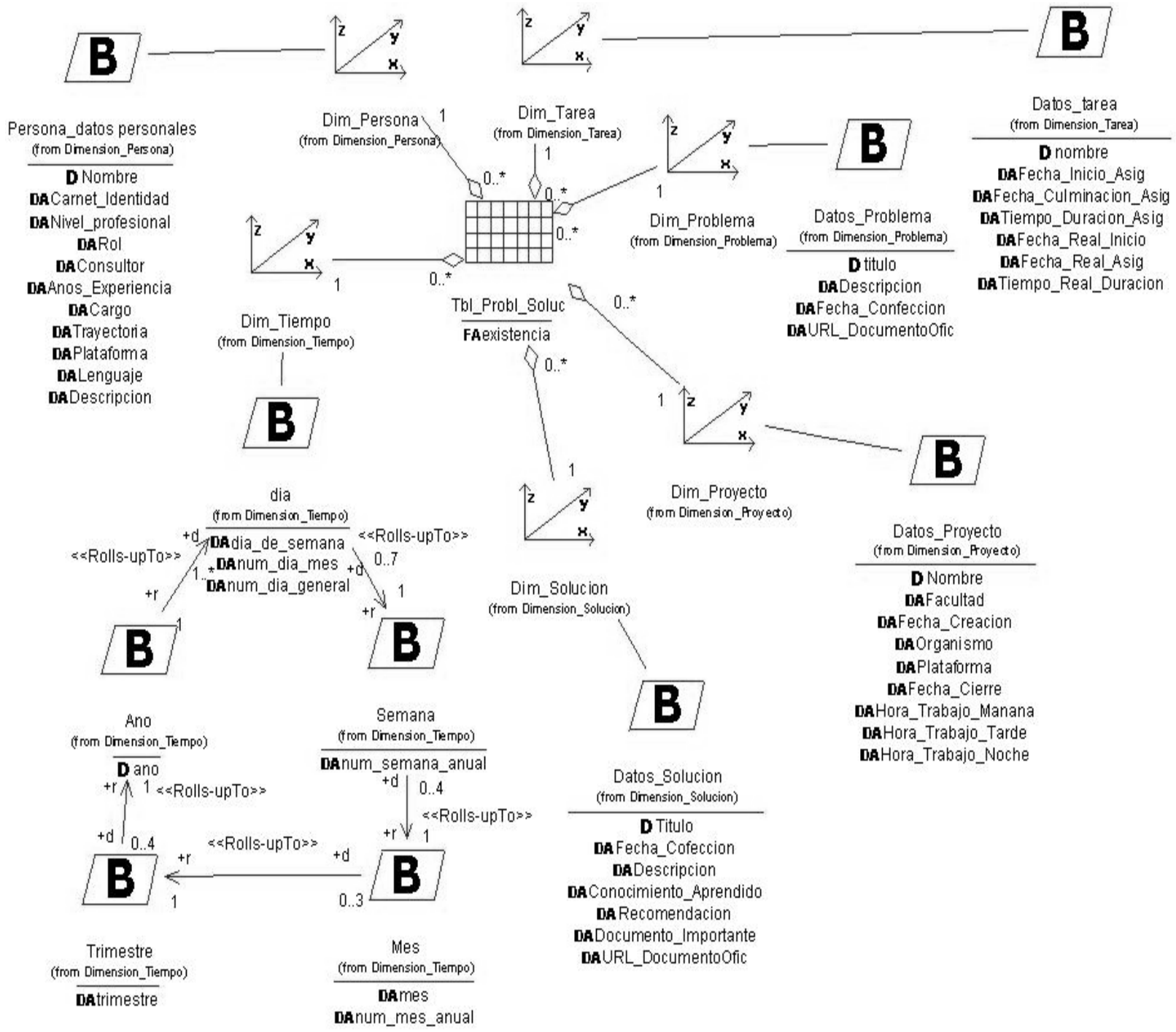


Figura 3.6 Esquema conceptual del DWH. Paquete hecho Pack Fact_Probl_Soluc

En el anexo 3.6 se representa lo incluido en el paquete Schema_Calif_Probl_Soluc (nivel 2) con sus respectivos paquetes hecho y dimensiones. El nivel 3 del paquete hecho se encuentra en el anexo 3.7 así como el de los paquetes dimensiones: Dimension_Probl_Sin_CGeneralizada y Dimension_Soluc_Sin_CGeneralizada en los anexos 3.8 y 3.9 respectivamente pues el resto de los paquetes dimensiones han sido importadas del paquete estrella Schema_Probl_Soluc.

El nivel 2 del paquete estrella Schema_Historial_Persona se encuentra en el anexo 3.10 con su paquete hecho Pack Fact_Historial_Persona en el medio y sus dimensiones, las cuales pertenecen al paquete Schema_Probl_Soluc. El contenido del paquete hecho (nivel 3) se puede consultar en el anexo 3.11.

Para tener una visión general de todas las tablas hechos, dimensiones y dependencias que contiene el GestCon Mart puede remitirse al anexo 3.12, en el cual todas las definiciones de los esquemas estrellas se han asociado en un único diagrama (nivel2).

3.4.3.2 Mapeo de datos

El proceso de consolidación de la información en el DWH involucra actividades de extracción de diversas fuentes de datos, transformación y finalmente su carga en el DWH, a este proceso se le denomina ETL.

Las transformaciones aplicadas a los datos provenientes de las distintas fuentes son básicamente de limpieza y de estructuración. Las transformaciones de limpieza son necesarias para asegurar la calidad de los datos finalmente almacenados en el DWH e incluye entre otros, la corrección de errores, eliminación de redundancia y resolución de inconsistencias. Los cambios en la estructura se realizan para adecuar los esquemas a las funcionalidades de un DWH, e incluyen la adecuación al modelo de datos del DWH, cambios de formato, operaciones de agregación, etc.

Para representar el flujo de datos desde las diversas fuentes hacia el DWH se utilizará el diagrama de mapeo de datos (Data Mapping), con varios niveles de detalle. Debido a que puede tornarse muy complejo, el autor (LUJAN-MORA 2005) en su propuesta lo divide en cuatro niveles:

- Nivel de base de datos o Nivel 0: En este nivel cada esquema del almacén de datos se representa mediante un paquete. Los mapeos entre los diferentes esquemas se modelan en un único paquete de mapeo, que encapsula todos los detalles.

- Nivel de flujo de datos o Nivel1: Este nivel describe las relaciones de datos a nivel individual entre las fuentes de datos hacia los respectivos destinos en el almacén de datos.
- Nivel de tabla o Nivel2: Mientras que el diagrama de mapeo en el nivel 1 describe las relaciones entre las fuentes y los destinos de datos mediante un único paquete, el diagrama de mapeo de datos en el nivel de tabla detalla todas las transformaciones intermedias que tienen lugar durante ese flujo.
- Nivel de atributo o Nivel 3: En este nivel, el diagrama de mapeo de datos captura los mapeos existentes a nivel de atributo.

En la figura 3.7 se puede consultar el nivel 0 del mapeo de datos representado por un paquete llamado Data Mapping relacionado con el esquema conceptual de la fuente (SCS) y el esquema conceptual del data warehouse (DWCS).

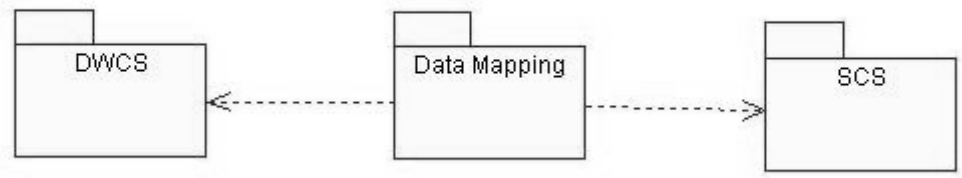


Figura 3.7 Mapeo de datos nivel 0

El GestCon Mart posee 11 tablas que se quieren poblar: 3 de hechos y 8 de dimensiones, por lo que existen 10 escenarios (nivel 1) dentro del paquete Data Mapping, uno para cada uno de las tablas a excepción de la tabla dimensión Tiempo que se llena junto con la dimensión Persona. Este nivel 1 representa el flujo de datos existente entre la fuente y el destino de datos en el contexto de cada escenario, descrito en la figura 3.8.

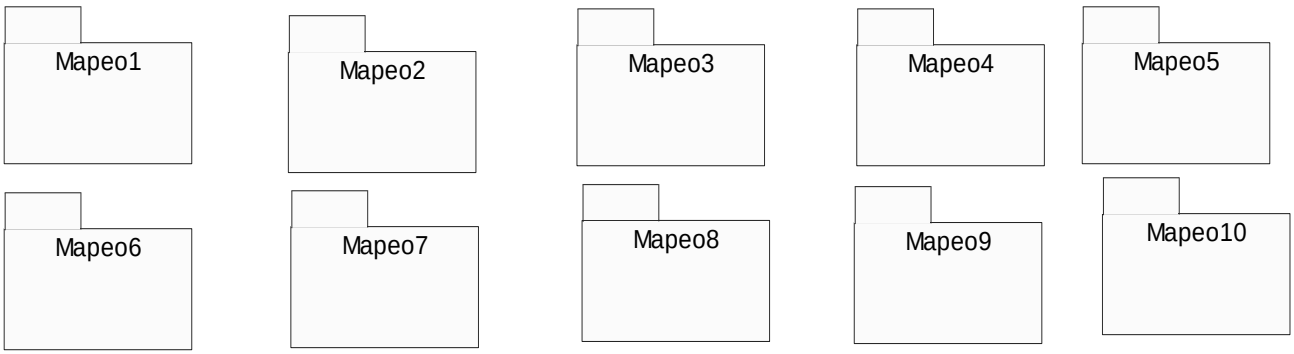


Figura 3.8 Mapeo de datos nivel 1

Para obtener una mejor visión del Mapeo de datos puede acudir a la figura 3.9 que muestra las peculiaridades del Mapeo1. En la misma los datos de la fuente (SCS) sufren 6 transformaciones (nivel2), es decir se transforman en 6 pasos. Existe un almacenamiento temporal denominado Intermedio que almacena los datos generados en los pasos 1, 2 y 3 los cuales se reciben como entrada en el paso 4 y también corresponde al almacenamiento temporal, luego en los pasos 5 y 6 se obtienen finalmente los valores deseados para las tablas del DWH Dimensión Persona y Dimensión Tiempo.

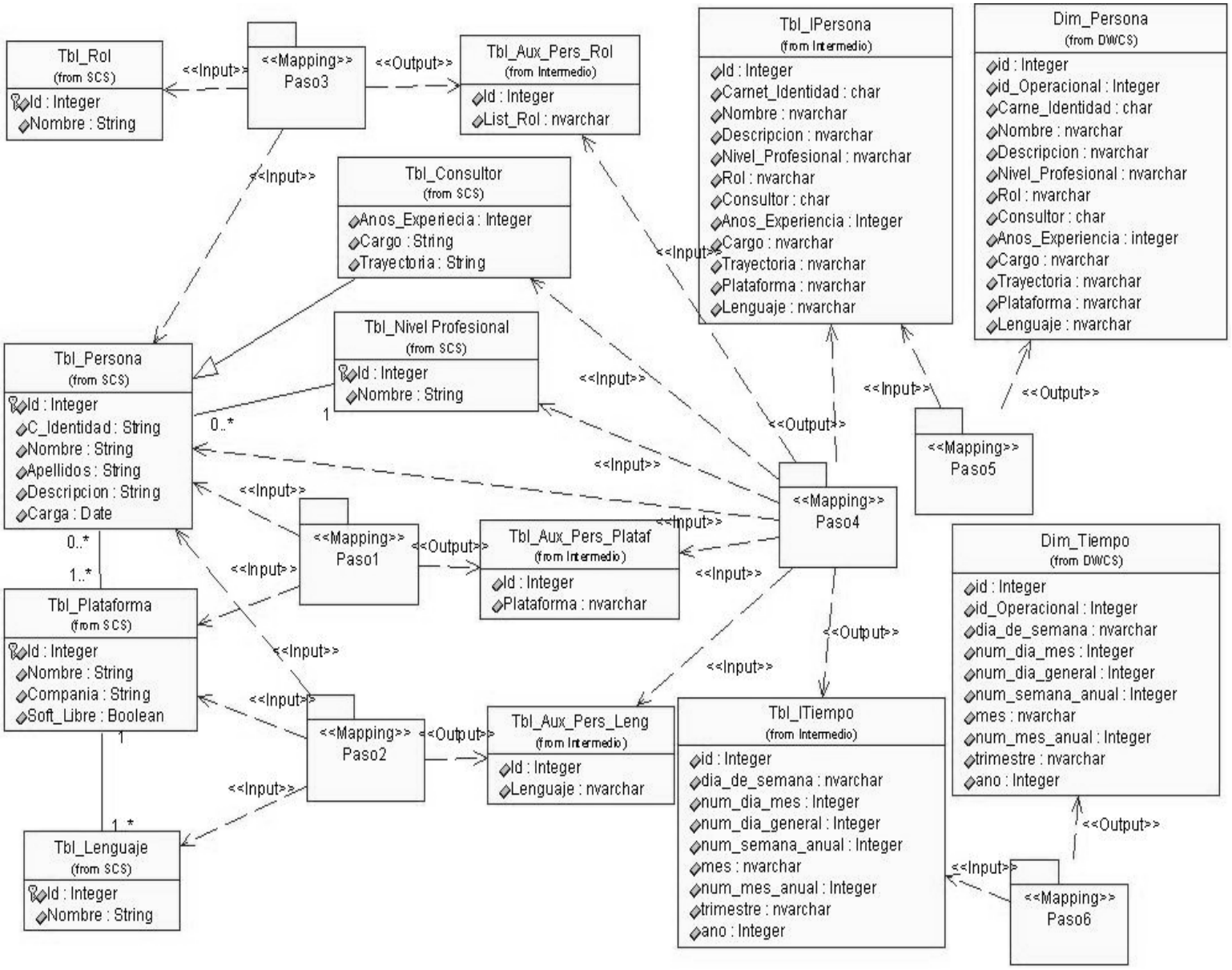


Figura 3.9 Mapeo1 (nivel2)

Asimismo en la figura 3.10 se puede apreciar cómo ocurre el mapeo a nivel de atributo (Nivel 3) del paso1 en el mapeo1.

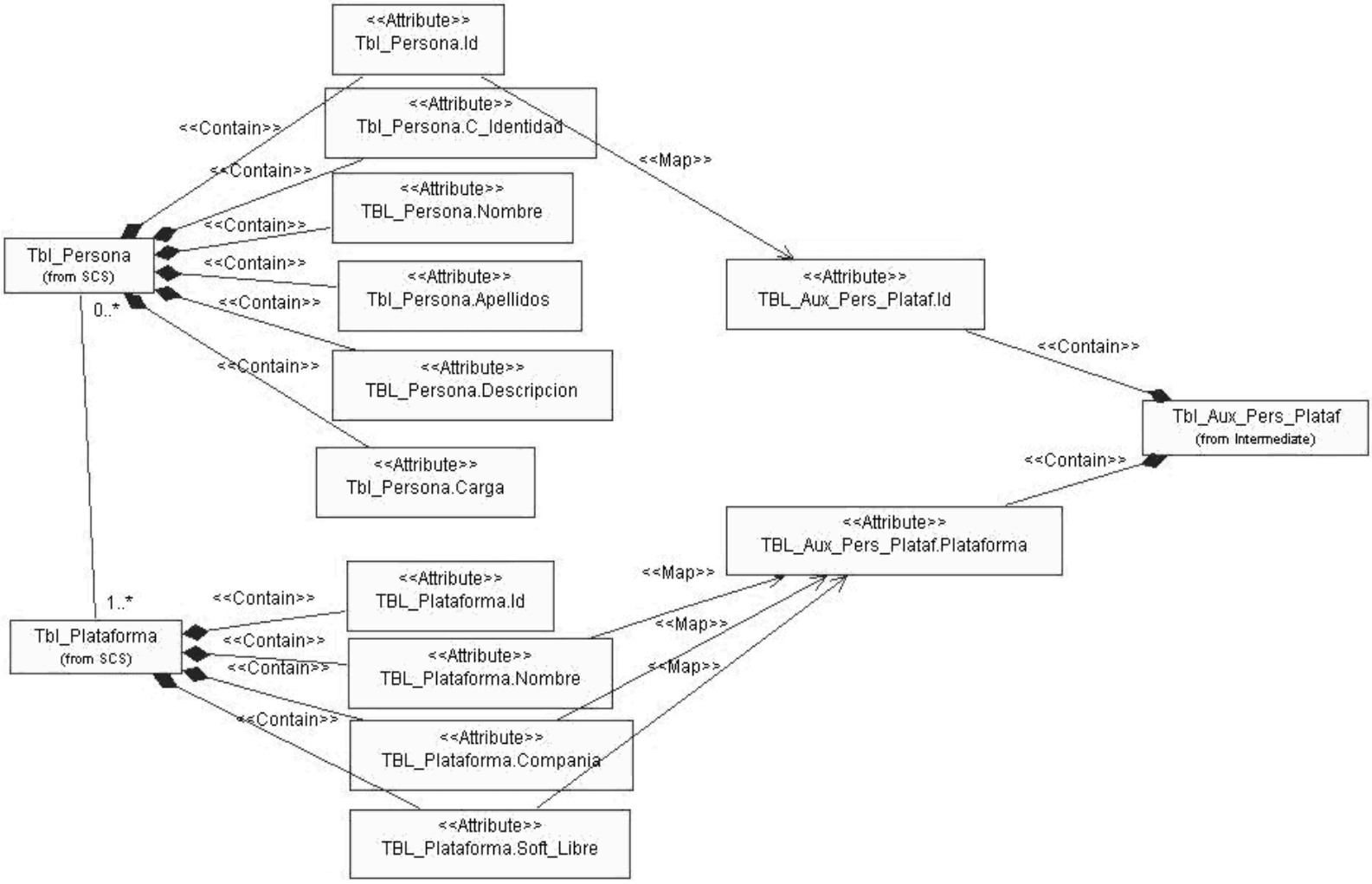


Figura 3.10 Mapeo1. Paso1 (nivel3)

Los restantes pasos de este mapeo1 se pueden consultar en los anexos 3.13, 3.14, 3.15, 3.16, 3.17.

En el anexo 3.18 se representa el nivel 2 del mapeo2 y su correspondiente mapeo a nivel de atributo en los anexos 3.19, 3.20 y 3.21.

En el anexo 3.22 se expone el nivel de tabla del mapeo3 así como su mapeo a nivel de atributo en los anexos 3.23, 3.24, 3.25.

También se presenta el nivel 2 del mapeo4 en el anexo 3.26 y el contenido de los diferentes pasos se puede consultar en los anexos 3.27, 3.28, 3.29, 3.30.

El nivel 2 del mapeo5 está en el anexo 3.31 y el mapeo a nivel de atributo en el anexo 3.32 y 3.33. El nivel de tabla del mapeo6 se puede encontrar en el anexo 3.34 y las transformaciones en el anexo 3.35. El mapeo7 se puede consultar en el anexo 3.36 así como sus transformaciones en el 3.37. Se muestra en el anexo 3.38 el mapeo8 y el contenido de los diferentes pasos realizados se pueden consultar en los anexos 3.39, 3.40, 3.41, 3.42. En el anexo 3.43 se representa el mapeo9 con su correspondiente mapeo del nivel 3 en el anexo 3.44 y finalmente el mapeo 10 se puede analizar en el anexo 3.45 con sus mapeos a nivel de atributo, en los anexos 3.46 y 3.47; en el paso1 de este último mapeo, el atributo id de las tablas Tbl_Proyecto, Tbl_Solucion, Tbl_Problema y Tbl_Persona que se mapea con Tbl_Relac_Probl_Soluc también se usa para buscar y actualizar el campo carga de cada una de estas tablas, con la fecha del día actual. En el paso 2 después que se modela el mapeo de datos para la tabla Tbl_Probl_Soluc, se eliminan todos los datos del área de reconciliación (Intermedio) y las tablas involucradas son: Tbl_Aux_Pers_Plataf, Tbl_Aux_Pers_Leng, Tbl_Aux_Pers_Rol, Tbl_Aux_Problema, Tbl_Aux_Probl_Relacion, Tbl_Aux1, Tbl_Aux_Proy_Plataf, Tbl_Aux_Solucion, Tbl_Aux_Solucion_Relacion, Tbl_Aux_Documento_Importante, Tbl_Aux, Tbl_Aux_Relacion_ListIdSoluciones, Tbl_Aux_Relacion_ListIdTareas, Tbl_Aux_Relacion_ListIdProblemas, Tbl_V_List_Pers_Proy_Problema, Tbl_V_List_Pers_Proy_Tarea, Tbl_V_List_Pers_Proy_Solucion, Tbl_Relacion, Tbl_Relacion_Probl_Soluc, Tbl_Proyecto, Tbl_Tiempo, Tbl_Tarea, Tbl_Solucion, Tbl_Problema, Tbl_Persona.

3.4.4 Implementación

Un diagrama de implementación muestra las dependencias entre las partes de código del sistema (diagrama de componentes) o la estructura del sistema en ejecución (diagrama de despliegue): estos muestran las relaciones físicas entre los componentes hardware y software en el sistema final, es decir, la configuración de los elementos de procesamiento en tiempo de ejecución y los componentes software (procesos y objetos que se ejecutan en ellos), mientras que un diagrama de componentes muestra las organizaciones y dependencias lógicas entre componentes software.

Durante este flujo de trabajo son construidas las estructuras físicas del DWH, es definido el tipo de almacenamiento empleado, se lleva a cabo el proceso de población de datos, se definen los procesos de exportación y se implementan los informes solicitados por los usuarios haciendo uso de la herramienta de consulta empleada, generalmente una aplicación OLAP.

Los principales diagramas de este flujo de trabajo son **el *esquema lógico del DWH,* el esquema físico de las fuentes (SPS), el *esquema físico del DWH*** y **el diagrama de transportación**.

3.4.4.1 Esquema lógico del DWH

En la figura 3.11 se representa uno de los cubos de datos que conforman al GestCon Mart, a nivel lógico, con sus respectivas tablas de hechos y dimensiones. En los anexos 3.48 y 3.49 se muestran los restantes cubos de datos.

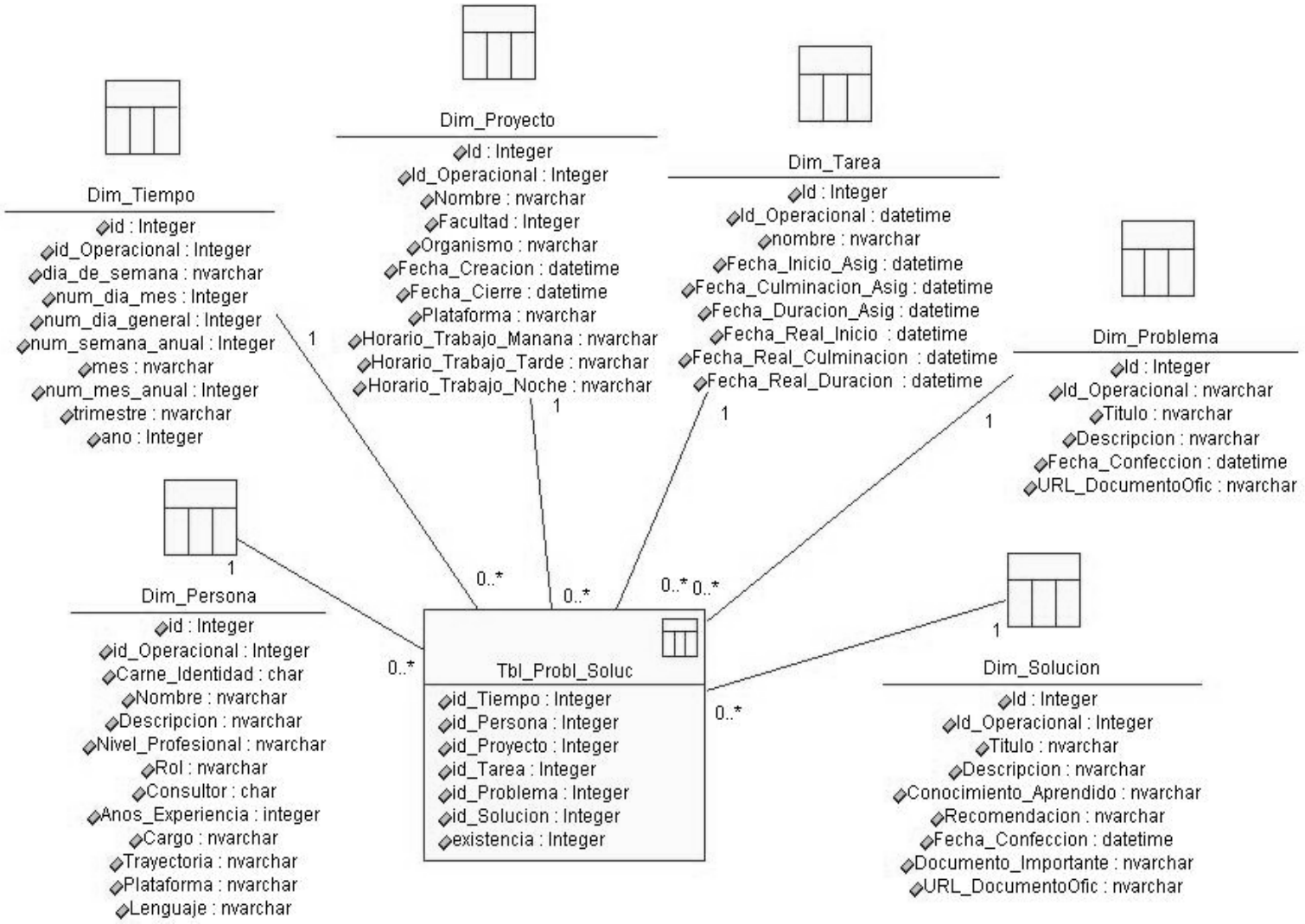


Figura 3.11 Esquema lógico de DWH. Cubo de dato Problemas y Soluciones.

3.4.4.2 Esquema Físico de la Fuente

Este esquema describe los orígenes de los datos del DWH desde un punto de vista físico. En la figura 3.12 se muestra el correspondiente diagrama, constituido por un servidor llamado KnowledgeServer dentro del cual se encuentra la base de datos operacional Knowledge. El nodo KnowledgeServer contiene una serie de valores que permiten describir las características particulares de dicho nodo.

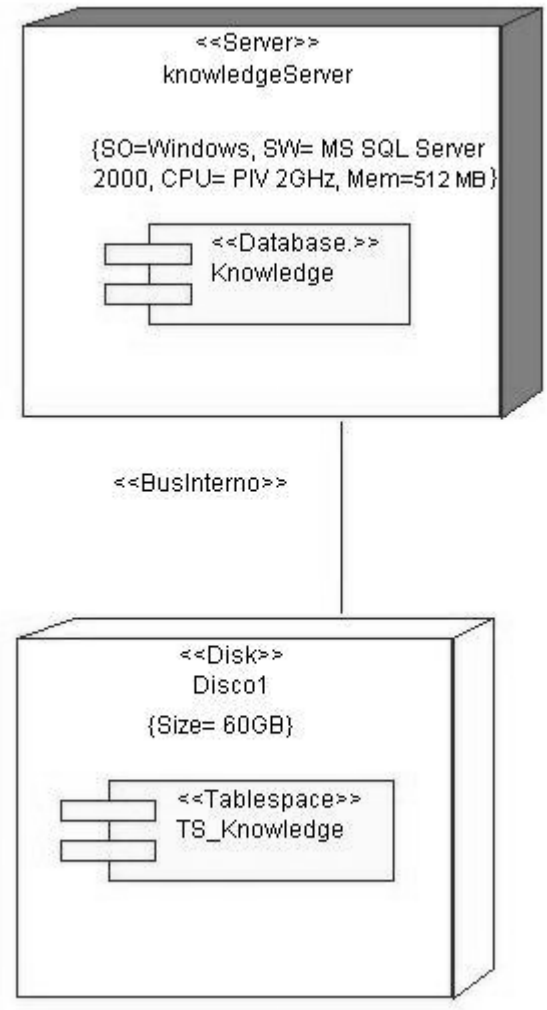


Figura 3.12 Esquema Físico de la Fuente

El estereotipo <<Server>> define una computadora que ejecuta funciones servidoras, <<Disk>> es utilizado para representar un disco de la PC, <<BusInterno>> para especificar el tipo de comunicación entre dos elementos y <<Tablespace>> representa la unidad lógica de almacenamiento de los ficheros físicos de la base de datos operacional Knowledge.

3.4.4.3 Esquema Físico del DWH

Este esquema muestra los aspectos físicos de la implementación del DWH, el mismo es dividido en dos partes, el diagrama de componentes y el diagrama de despliegue. En la figura 3.13 se puede observar que el DM Knowledge_DMart está formada por dos tablespace: Hechos y Diemensiones, la primera alberga a las tablas Tbl_Calif_Probl_Soluc, Tbl_Historial_Persona y Tbl_Probl_Soluc, y la segunda a la tablas Dim_Persona, Dim_Tiempo, Dim_Probl_Sin_CGeneralizada, Dim_Soluc_Sin_CGeneralizada, Dim_Tarea, Dim_Problema, Dim_Solucion y Dim_Proyecto.

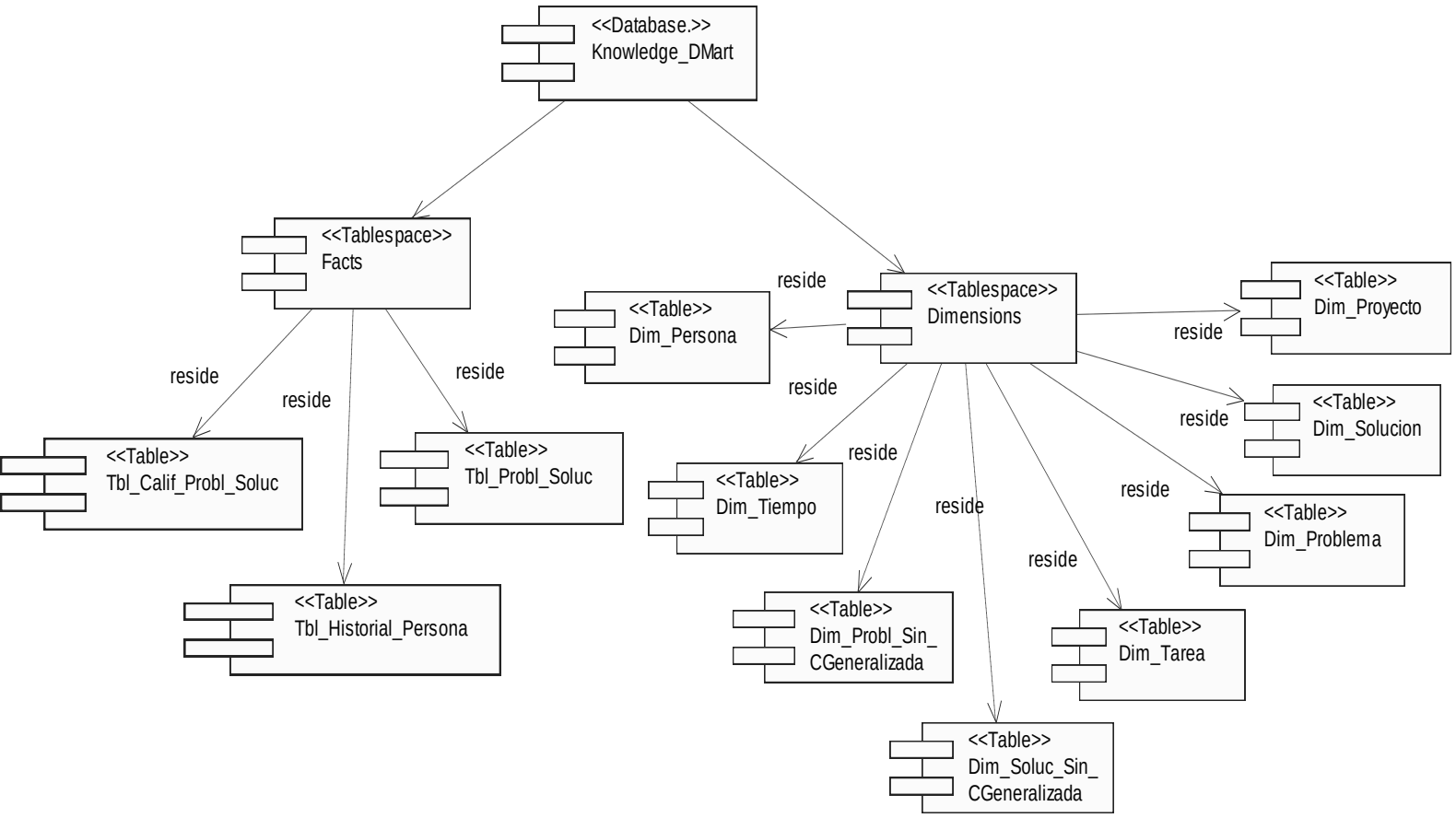


Figura 3.13 Esquema Físico del DWH. Diagrama de componentes

En el diagrama de despliegue son especificados los diferentes aspectos relativos a la configuración del software y del hardware, por otro lado la distribución de las estructuras lógicas definidas previamente también son representadas. En la figura 3.14 se puede observar la configuración del servidor que alberga al DM. El disco1 hospeda a las tablas hechos y dimensiones aunque es más recomendable utilizar discos diferentes para cada tablespace.

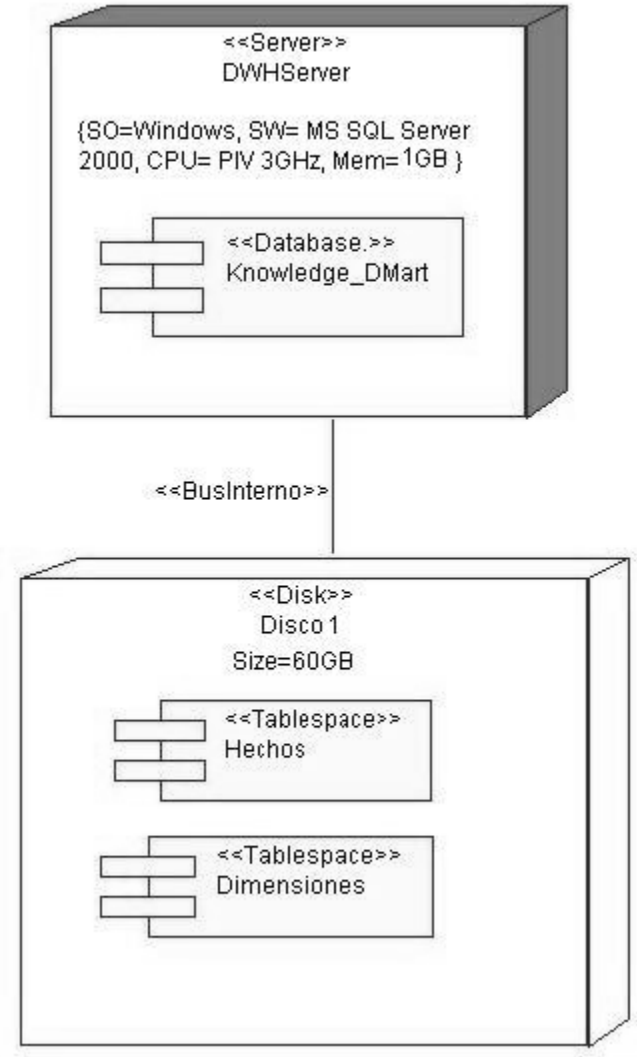


Figura 3.14 Esquema Físico del DWH. Diagrama de despliegue

3.4.4.4 Diagrama de integración de transporte

El diagrama de integración de transporte define la estructura física de los procesos ETL usados en la carga de datos para el DWH desde el origen de datos.

En la figura 3.15 se muestran dos servidores: knowledgeServer y DWHServer los cuales han sido definidos anteriormente, a este último se le agrega la tarea de ser también servidor de los procesos ETL, es decir el mismo servidor va a contener al DM y va a soportar los procesos de transformación de datos,

aunque lo más recomendable es que se introduzca un nuevo servidor para la ejecución de los procesos ETL. Dicho servidor se comunica con el knowledgeServer a través del protocolo OLEDB (Object Linking and Embedding DataBase) porque el software utilizado es Microsoft SQL Server 2000 para el DM y para los procesos de transformación el Data Transformation Services.

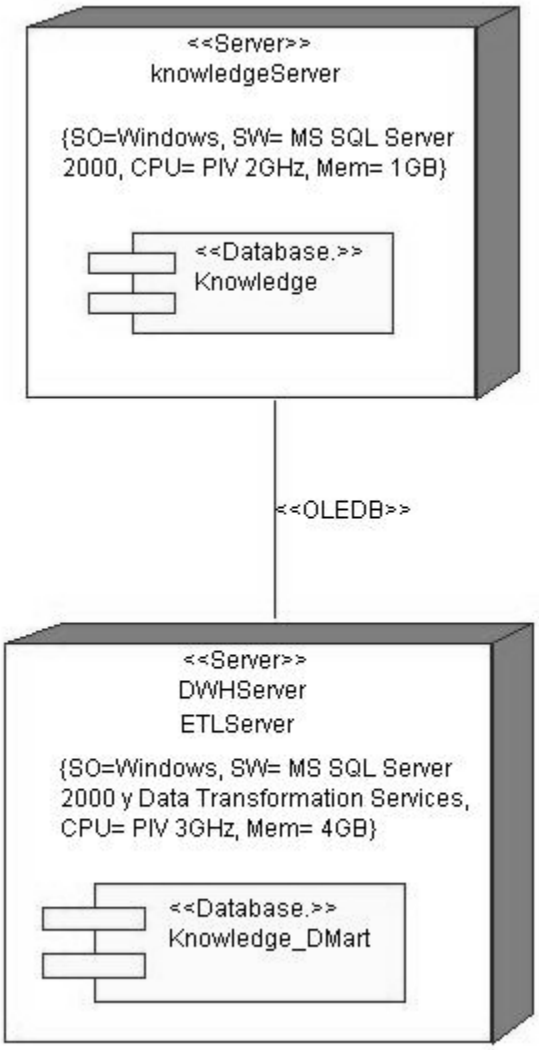


Figura 3.15 Diagrama de integración de transporte

3.4.5 Prueba

En este flujo de trabajo se verifica que la aplicación trabaja como se desea y tiene como propósito diseñar e implementar las pruebas, creando los casos de prueba, poniéndolas en práctica y analizando los resultados obtenidos.

3.4.6 Mantenimiento

El propósito de este flujo es definir el refrescamiento y los procesos de carga necesarios para mantener el DWH actualizado. Se Inicia cuando la construcción del almacén está terminada y es entregado al usuario final. No tiene una fecha terminación y dura toda la vida del DWH.

3.4.7 Revisión Pos-Desarrollo

En este flujo se lleva a cabo una revisión y análisis del tiempo y esfuerzo empleado en cada fase del DWH, para haciendo uso de dicha información se puedan mejorar los futuros trabajos.

3.5 Conclusiones

En este capítulo se realizó una descripción detallada del Gestcon Mart, exponiendo principalmente la información más relevante que va a almacenar, se explica además los diferentes enfoques para efectuar la exploración de datos. Queda expuesto el proceso de diseño de un Data Warehouse y la aplicación del método DWEP para la construcción del Data Mart, haciendo uso de diferentes diagramas realizados en tres niveles: conceptual, lógico y físico.

4 Conclusiones

La gestión de conocimiento es una tarea imprescindible para una organización que aspira a tener éxito en el mercado actual. Esta labor cobra una relevancia extraordinaria en las condiciones de desarrollo de la UCI debido a dos características esenciales:

- El 90 % de las actividades de desarrollo de software son llevadas a cabo por trabajadores del conocimiento.
- La característica de universidad productiva implica una rápida rotación de los trabajadores y por ello es necesario que los nuevos miembros tengan un mecanismo para adquirir el conocimiento necesario para desarrollar las tareas que se le asignen de manera eficaz en la menor brevedad posible.

En el presente trabajo se implementó un Data Mart (GestCon Mart) que es capaz de dar soporte a una base de experiencia para la gestión de proyectos de software. Con lo cual se garantiza el rápido acceso por parte de los líderes de proyectos a las experiencias acumuladas de la organización.

Al mismo tiempo es capaz de servir de repositorio de datos para la aplicación de técnicas de minería de datos e inteligencia artificial que posibilitarían extraer el conocimiento tácito almacenado en los datos y convertirlo en conocimiento explícito con valor de uso para la organización.

A lo largo del trabajo de tesis se definieron y materializaron los artefactos propuestos por la metodología DWEP para la implementación de Data Warehouse, el uso de esta metodología permite que el Data Mart construido sea lo suficientemente flexible como para que pueda ser extendido en el futuro a otros procesos de la organización (Ej. Ingeniería de software).

5 Recomendaciones

Desarrollar trabajos para la aceleración de Data Marts enfocados a la cadena valor de la universidad, orientados al almacenamiento de los datos referente a la formación académica y a la parte ingenieril para mejores prácticas de ingeniería de software, con vistas a que se extienda la base de experiencias construida como parte de este trabajo. De forma tal que pueda centralizarse de manera automática todo el trabajo de gestión de conocimiento.

Abrir líneas de investigación cuyos objetivos sean la aplicación de algoritmos de minería de datos a la base existente de manera que contribuyan a la transformación del conocimiento tácito a conocimiento explícito.

6 Referencias Bibliográficas

ABELLÓ, A.; J. SAMOS*, et al. YAM^2 (Yet Another Multidimensional Model: An Extension of UML*. Proceedings of the International Database Engineering and Application Symposium, Edmonton, Canada, IEEE Computer Society, 2002. 172-181 p.

BATINI, C.; S. CERI*, et al. Conceptual Database Design: An Entity-Relationship Approach*. Benjamin-Cummings Publishing Company 1992.

BLASCHKA, M.; C. SAPIA*, et al. Extending the E/R Model for the Multidimensional Paradigm*. Proceedings of the 1st International Workshop on Data Warehouse and Data Mining, Singapore, Springer-Verlag, 1998. 105-116 p.

CABIBBO, L. and R. TORLONE. *Advances in Database Technology*. 6th International Conference on Extending Database Technology, Valencia, Spain, Springer-Verlag, 1998. 183-197 p.

CODD, E. F.; S. B. CODD*, et al. Providing OLAP to User-Analysts: An IT Mandate*. E.F. Codd Associates, 1993.

DALGLEISH, D. *Excel Pivot Tables Recipe Book: A Problem-Solution Approach*. USA, Apress, 2006. 336 p. 1590596293

ECKERSON, W. Four Ways to Build a Data Warehouse *Application Development Trends*, 2002.

GOLFARELLI, M. and S. RIZZI. *A Methodological Framework for Data Warehouse Design*. Proceedings of the ACM 1st International Workshop on Data Warehousing and OLAP, Bethesda, USA, ACM, 1998. 3-9 p.

HACKNEY, D. *Data Warehouse Delivery: Who Are You?* . *DM Review Magazine*, 1998. Part I.

INMON, W. *Building the Data Warehouse*. 3ra. John Wiley & Sons, Inc, 2002. 432 p. 0471081302

INMON, W. *Building the data warehouse*. 2da. New York, USA, John Wiley & Sons, Inc, 1996. 401 p.

INMON, W. H. *Building the data warehouse*. New York, USA, QED Information Sciences, Inc, 1992. 272 p. 0471569607

INMON, W. *Building the Data Warehouse*. 4ta. Indianapolis y Canada, Wiley Publishing, Inc., 2005. 543 p. 9780764599446

JACOBSON, I.; G. BOOCH*, et al. El Proceso Unificado de Desarrollo de Software*. Madrid, Addison-Wesley, 2000. 464 p.

JARKE, M.; M. LENZERINI*, et al. Fundamentals of Data Warehouses*. 2da. Springer-Verlag, 2003.

KIMBALL, R. *The Data Warehouse Toolkit*. 1a. New York, John Wiley & Sons, Inc, 1996. 388 p. 0471153370

KIMBALL, R. *Meta Meta Data Data*. *DBMS Magazine*, 1998. 11: 18-20 p.

KIMBALL, R. and J. CASERTA. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Indianapolis y Canadá, Wiley Publishing, Inc., 2004. 525 p. 0764567578

LIEBOWITZ, J. *Knowledge Management Handbook*. USA, CRC Press, 1999.

LUJAN-MORA, S. *Data Warehouse Design with UML*. España, Universidad de Alicante, 2005.

MARCO, D. *Building and Managing the Meta Data Repository: A Full Lifecycle Guide*. Wiley, 2000. 416 p. 9780471355236

MICROSOFT. *Libros en pantalla de SQL Server 2000*. Microsoft Corporation, 2000.

MICROSOFT. *Introducción a los Libros en pantalla de SQL Server 2005*. *Msdn*, Microsoft Corporation, 2007. http://msdn2.microsoft.com/es-es/library/ms166019(SQL.90).aspx

OMG. *OMG Unified Modeling Language Specification Version 1.5*. Massachusetts, Object Management Group, Inc, 2003. 736 p.

ORACLE. *10 g Oracle by Example - Oracle Developer Suite*. *TECHNOLOGY NETWORK*, 2007. http://www.oracle.com/technology/obe/obe_bi/index.html

PONNIAH, P. *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals.* New York , Chichester , Weinheim ,Brisbane , Singapore y Toronto, John Wiley & Sons, Inc., 2001. 518 p. 0471412546

SACHDEVA, S. *Meta Data Architecture for Data Warehousing*. *DM Review Magazine*, 1998.

SCHREIBER, G.; H. AKKERMANS*, et al. Knowledge engineering and management: the CommonKADS methodology*. USA MIT Press Cambridge, MA, 2000. 455 p. 0262193000

TANG, Z. and J. MACLENNAN. *Data Mining with SQL Server 2005*. Indianápolis, Wiley Publishing, Inc, 2005. 460 p. 978-0-471-46261-3

TRUJILLO, J.; M. PALOMAR*, et al. Designing Data Warehouses with OO Conceptual Models*. USA, IEEE Computer Society Press 2001. 66-75 p.

TRUJILLO, J.; M. PALOMAR*, et al. Extending UML for Multidimensional Modeling*. Proceedings of the 5th International Conference on the Unified Modeling Language, Dresden, Germany, Springer- Verlag, 2002. 290-304 p.

TRYFONA, N; F. BUSBORG*, et al. starER: A Conceptual Model for Data Warehouse Design*. Proceeding of ACM 2nd International Workshop on Data Warehousing and OLAP, Kansas City, USA, 1999. 3-8 p.

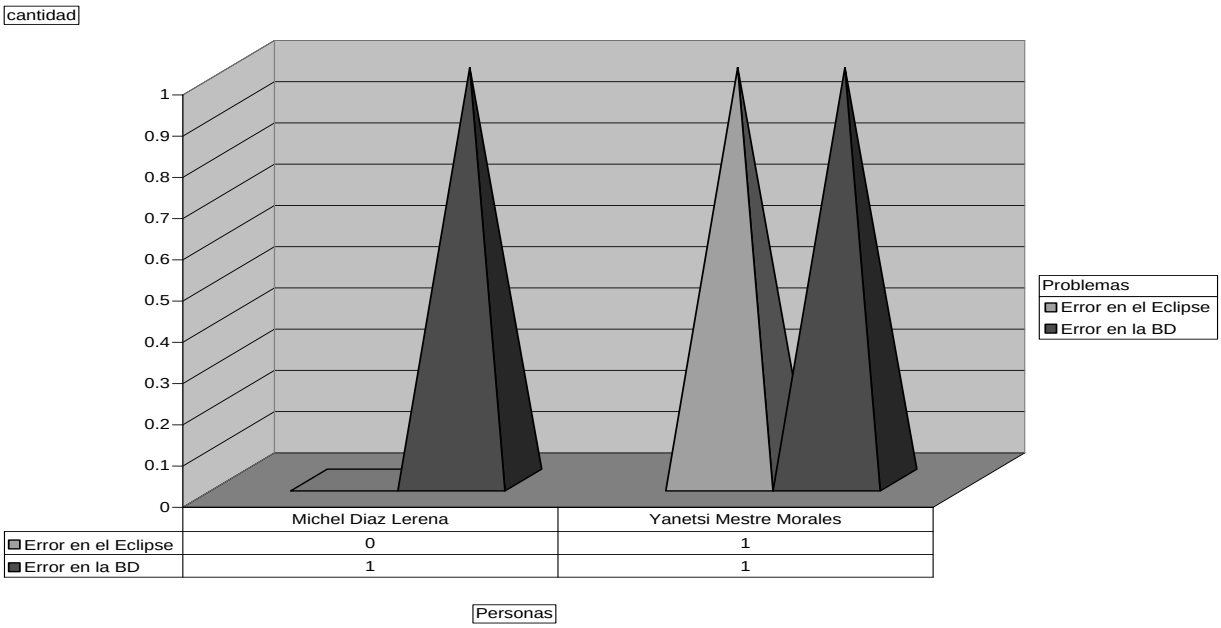
VIDAL, L. V. and M. V. MONTEAGUDO. *Estudio Teórico-Conceptual sobre Data Warehouse*. Ciudad de La Habana, Universidad de La Habana, Facultad de Matemática y Computación, 2000. 128 p.

7 Anexos

Anexo 3.1 Exploración en términos de personas y problemas

	A	B	C	D
1	Historial de Personas			
2				
3	cantidad	Problemas		
4	Personas	Error en el Eclipse	Error en la BD	Total
5	Michel Diaz Lerena	0	1	1
6	Yanetsi Mestre Morales	1	1	2
7	Total	1	2	3

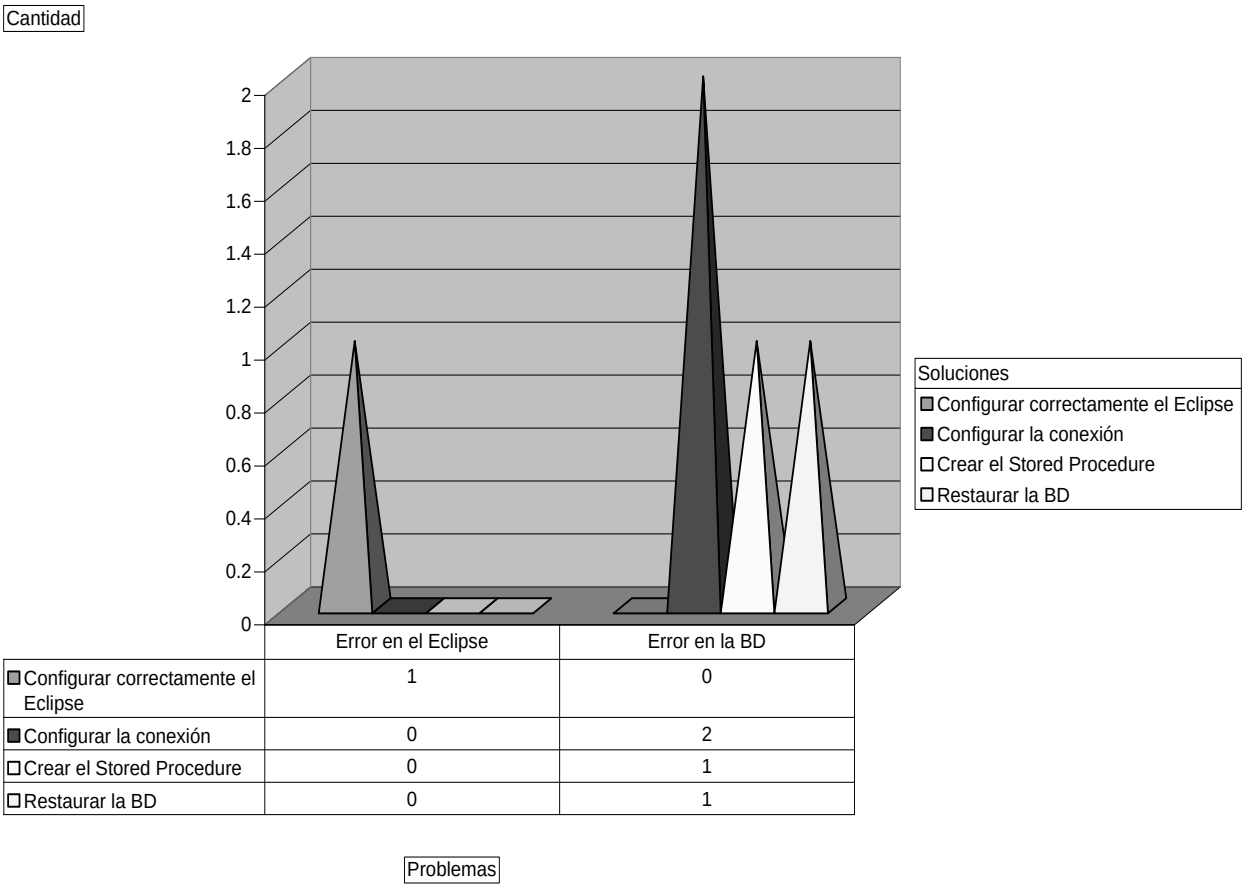
Anexo 3.2 Gráfico de la exploración en términos de personas y problemas



Anexo 3.3 Exploración en términos de problemas y soluciones

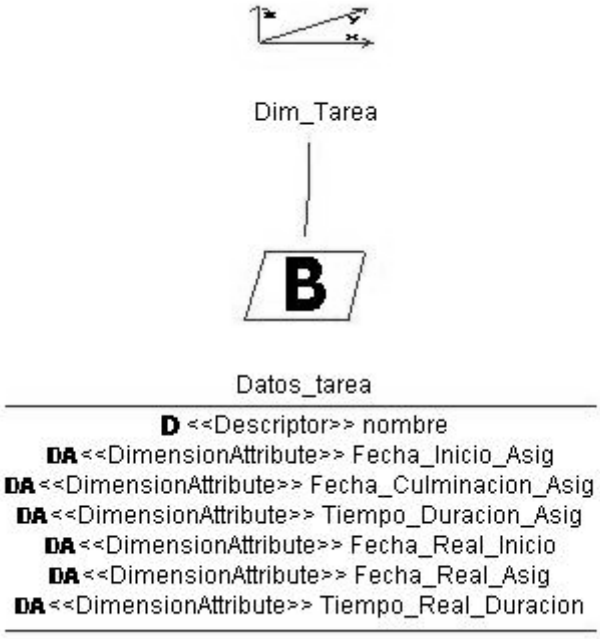
	A	B	C	D	E	F
1	Problemas y Soluciones					
2						
3	Cantidad	Soluciones				
4	Problemas	Configurar correctamente el Eclipse	Configurar la conexión	Crear el Stored Procedure	Restaurar la BD	Total
5	Error en el Eclipse	1	0	0	0	1
6	Error en la BD	0	2	1	1	4
7	Total	1	2	1	1	5

Anexo 3.4 Gráfico de la exploración en términos de problemas y soluciones.

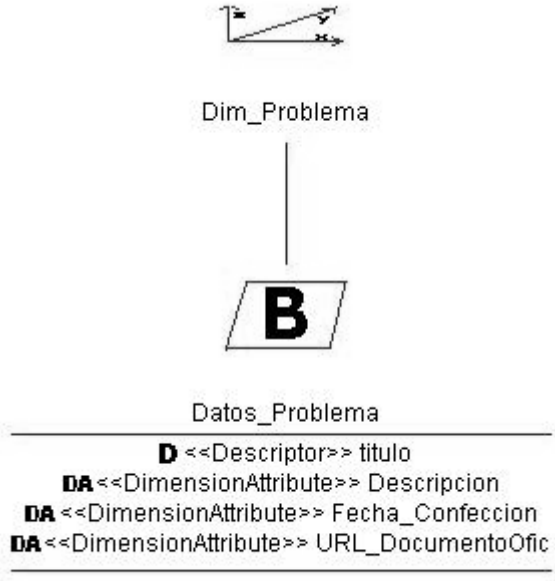


	Error en el Eclipse	Error en la BD
Configurar correctamente el Eclipse	1	0
Configurar la conexión	0	2
Crear el Stored Procedure	0	1
Restaurar la BD	0	1

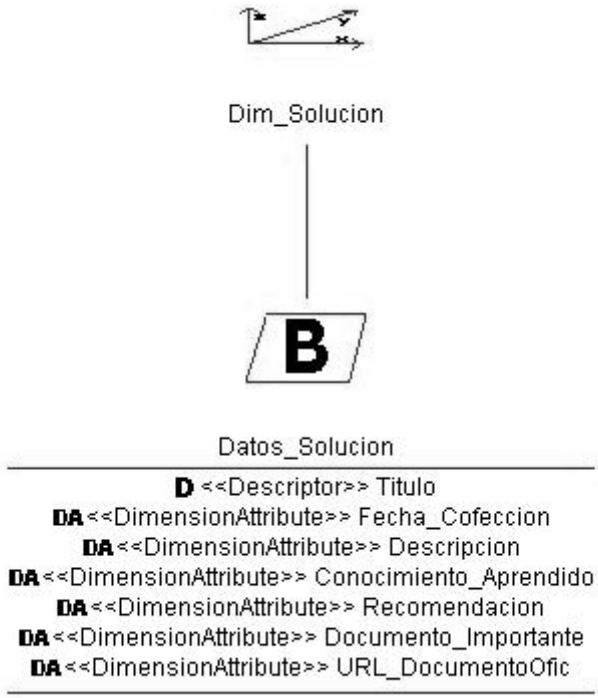
Anexo 3.5 Esquema conceptual del DWH. Dimension_Tarea (nivel 3)



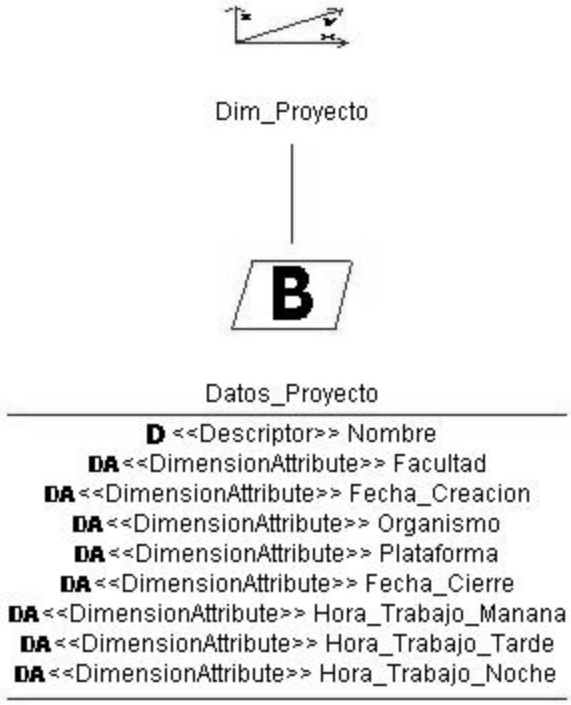
Anexo 3.6 Esquema conceptual del DWH. Dimension_Problema (nivel 3)



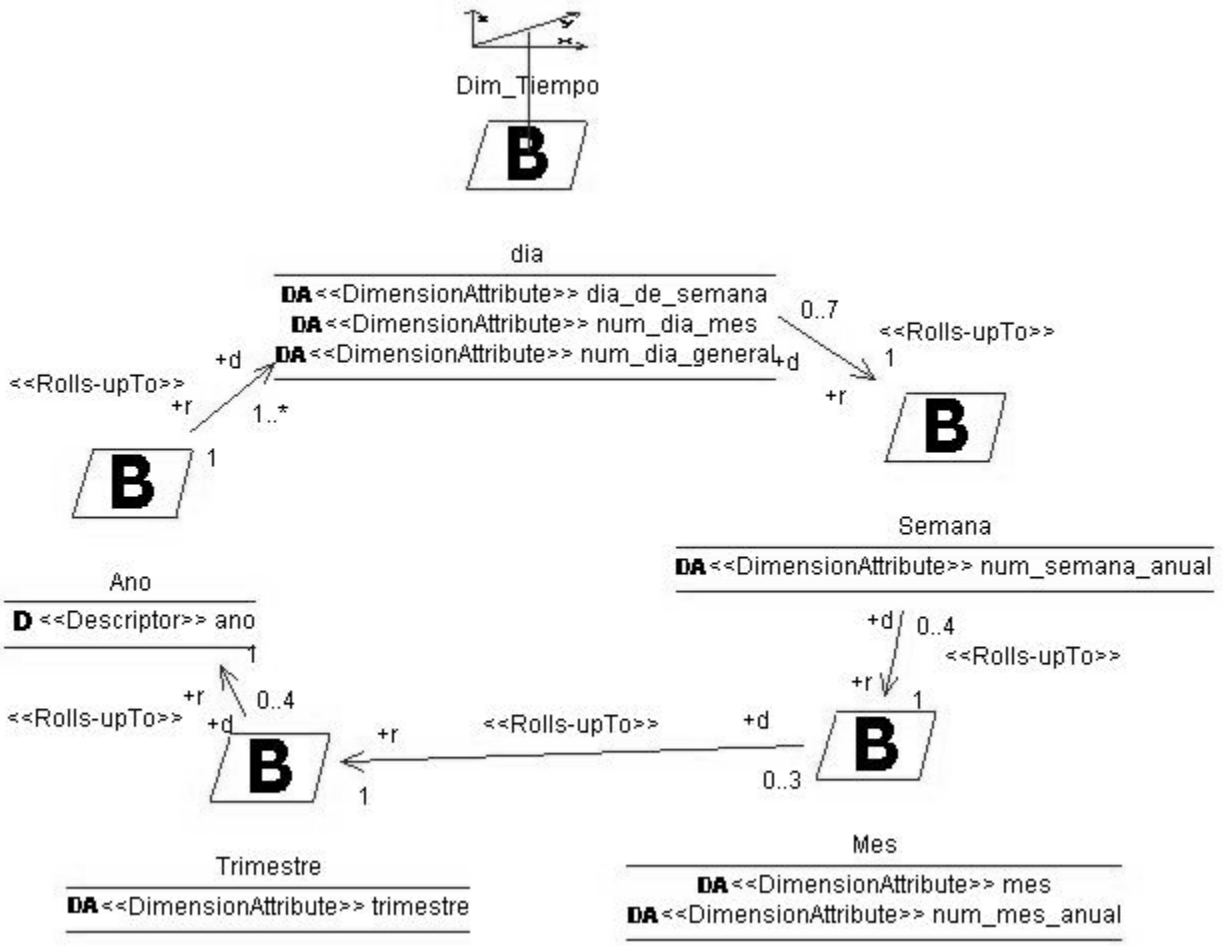
Anexo 3.7 Esquema conceptual del DWH. Dimension_Solucion (nivel 3)



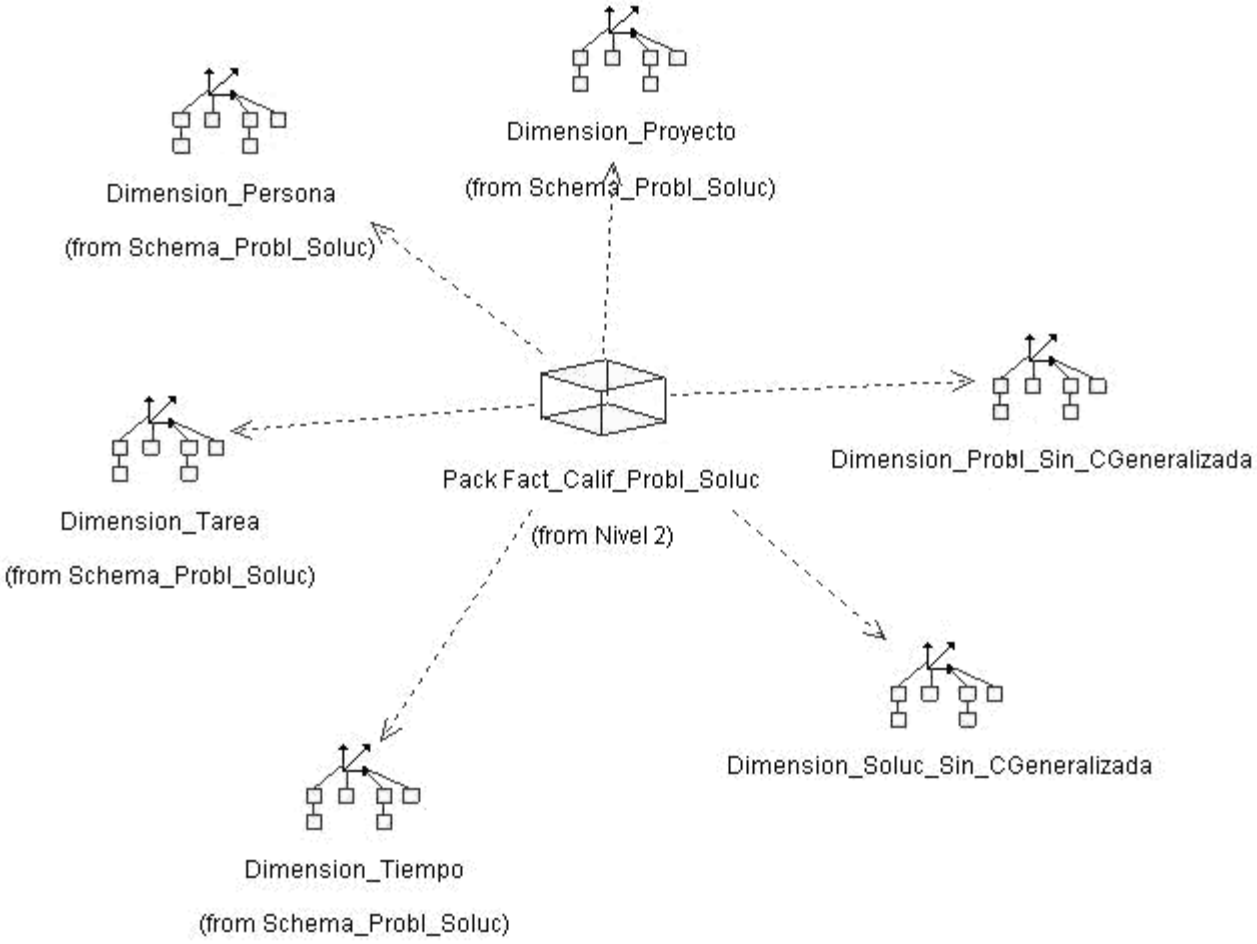
Anexo 3.8 Esquema conceptual del DWH. Dimension_Proyecto (nivel 3)



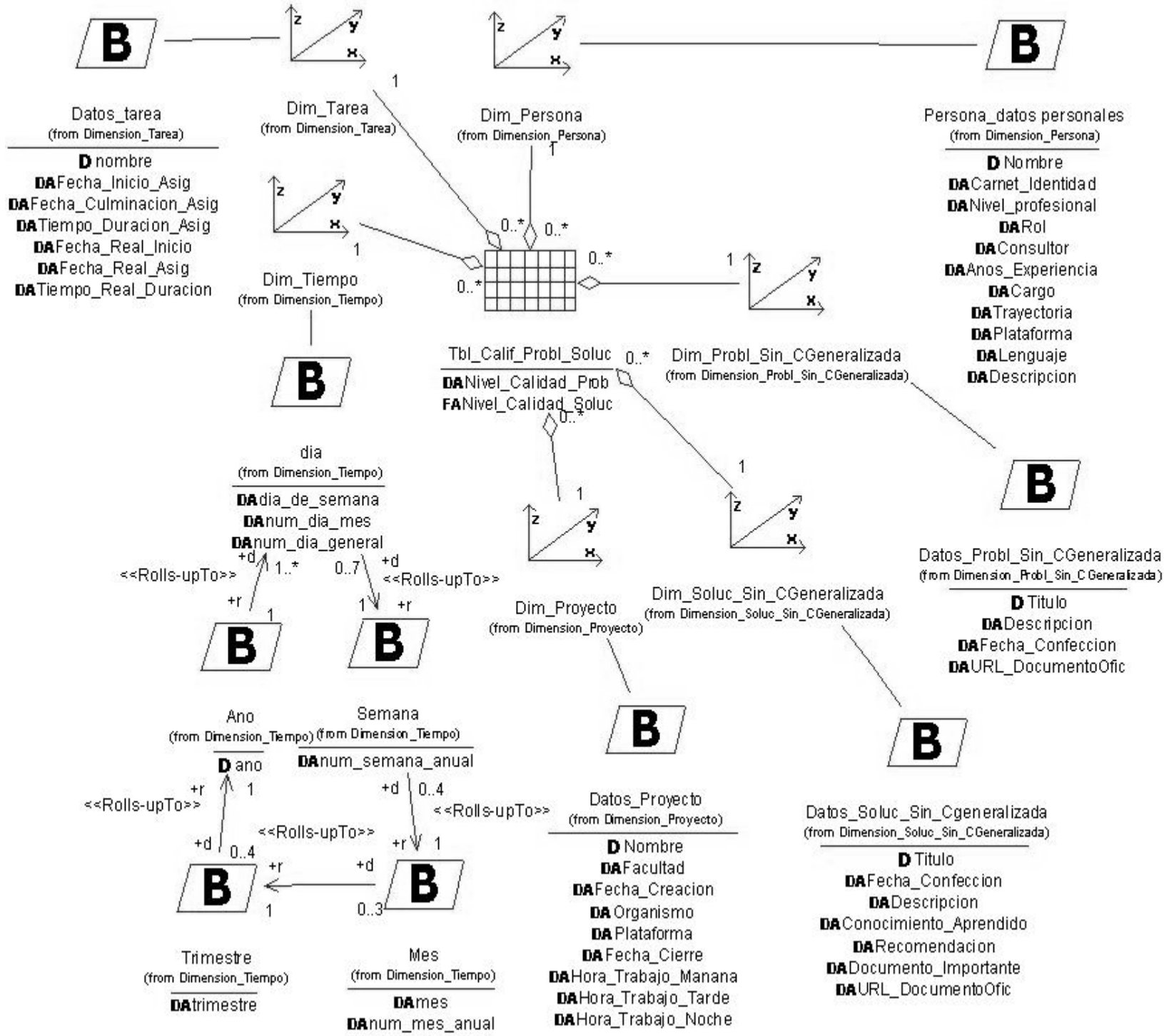
Anexo 3.9 Esquema conceptual del DWH. Dimension_Tiempo (nivel 3)



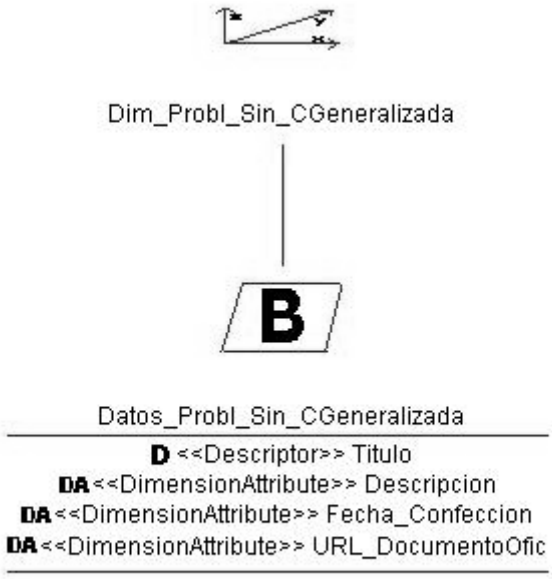
Anexo 3.10 Esquema conceptual del DWH. Schema_Calif_Probl_Soluc (nivel 2)



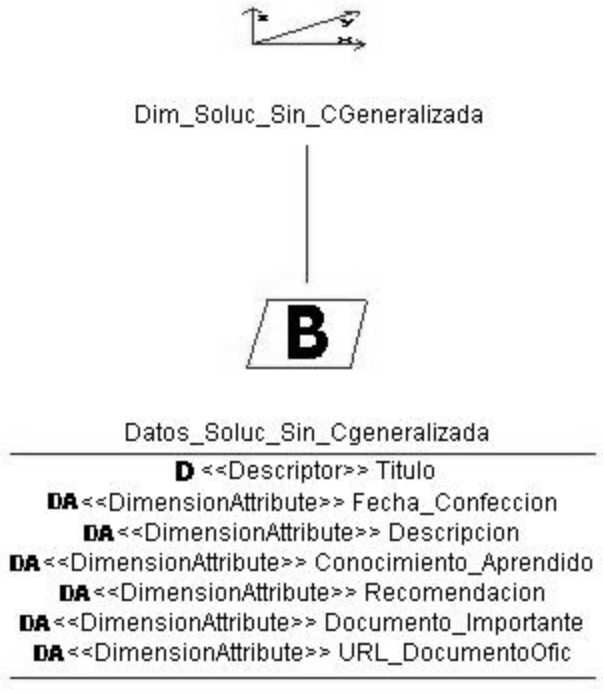
Anexo 3.11 Esquema conceptual del DWH. Paquete hecho Pack Fact_Calif_Probl_Soluc (nivel 3)



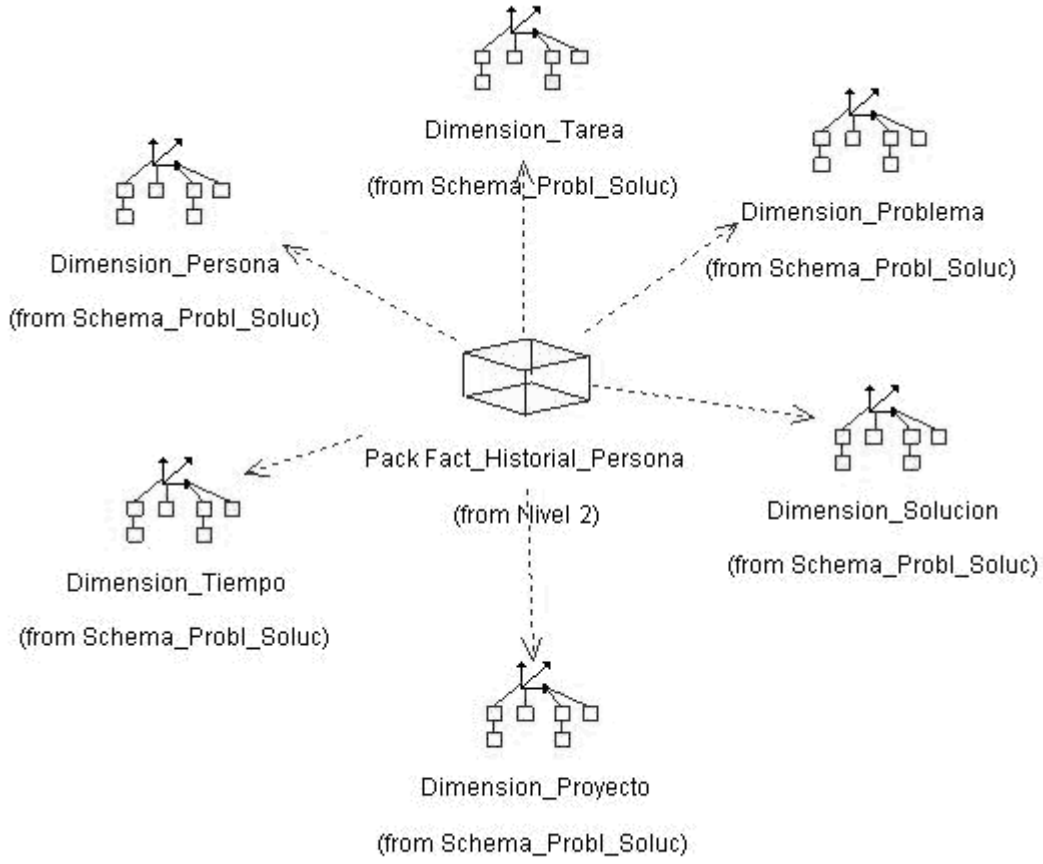
Anexo 3.12 Esquema conceptual del DWH. Dimension_Probl_Sin_CGeneralizada (nivel3)



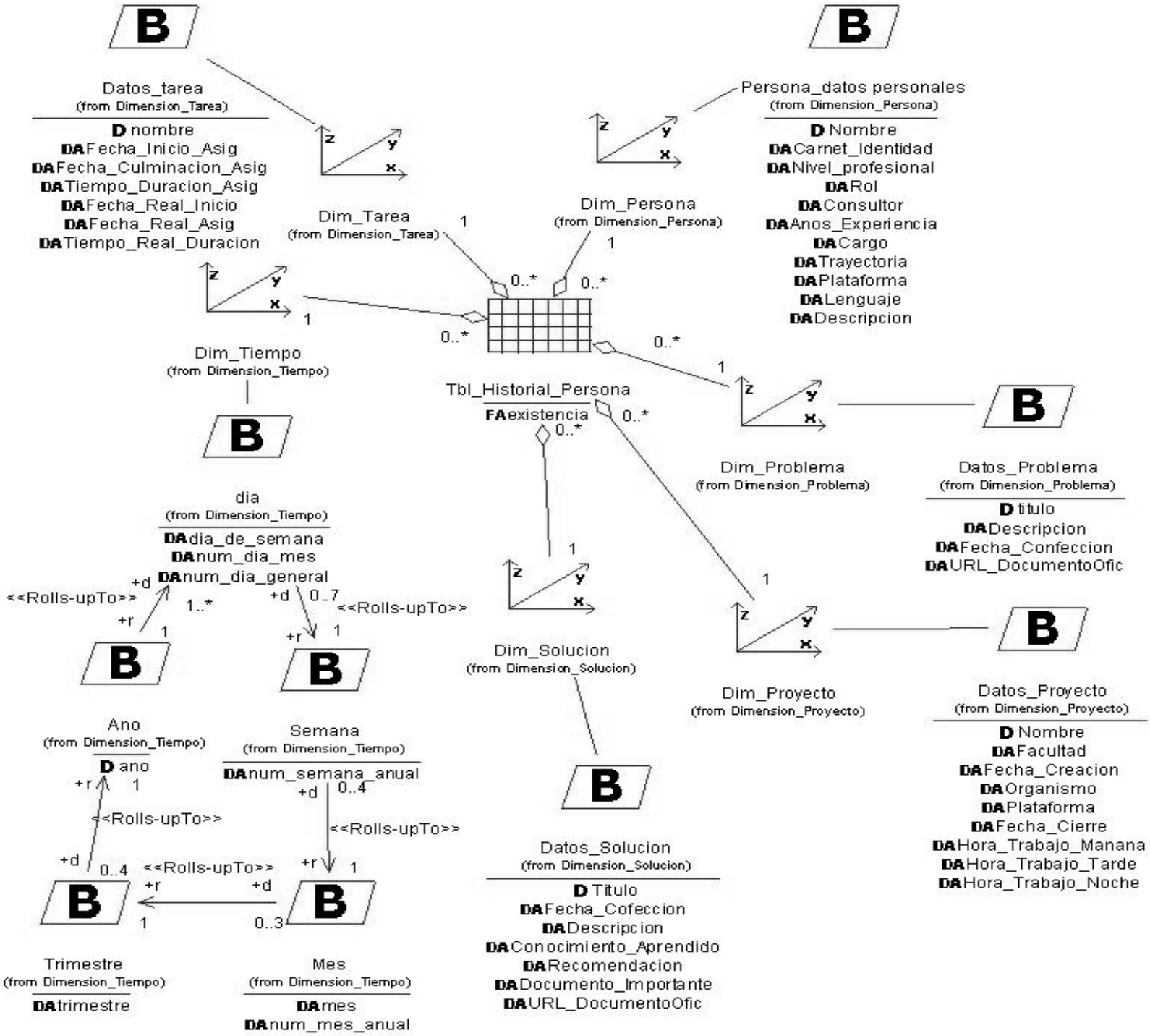
Anexo 3.13 Esquema conceptual del DWH. Dimension_Soluc_Sin_CGeneralizada(nivel 3)



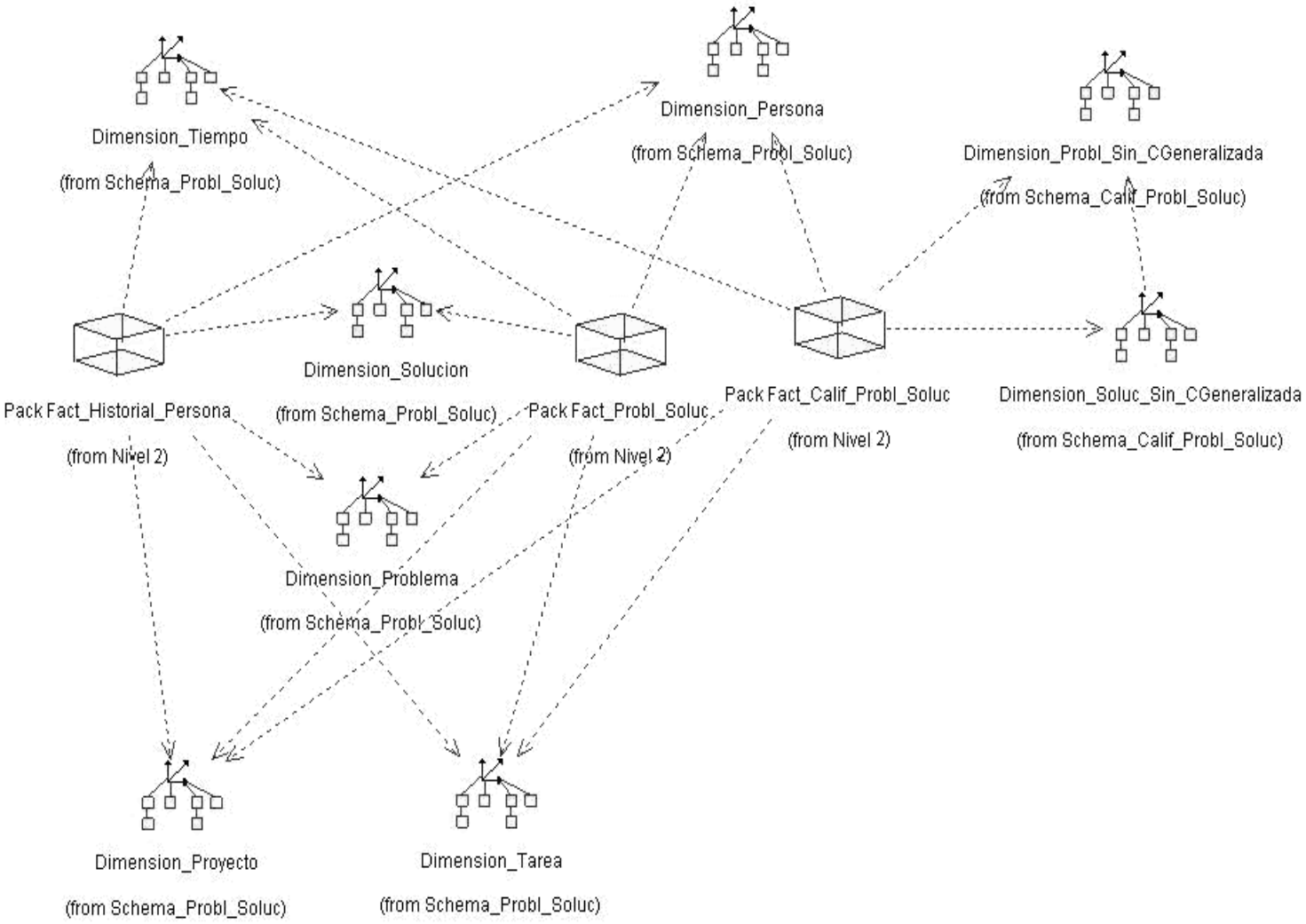
Anexo 3.14 Esquema conceptual del DWH. Schema_Historial_Persona (nivel 2)



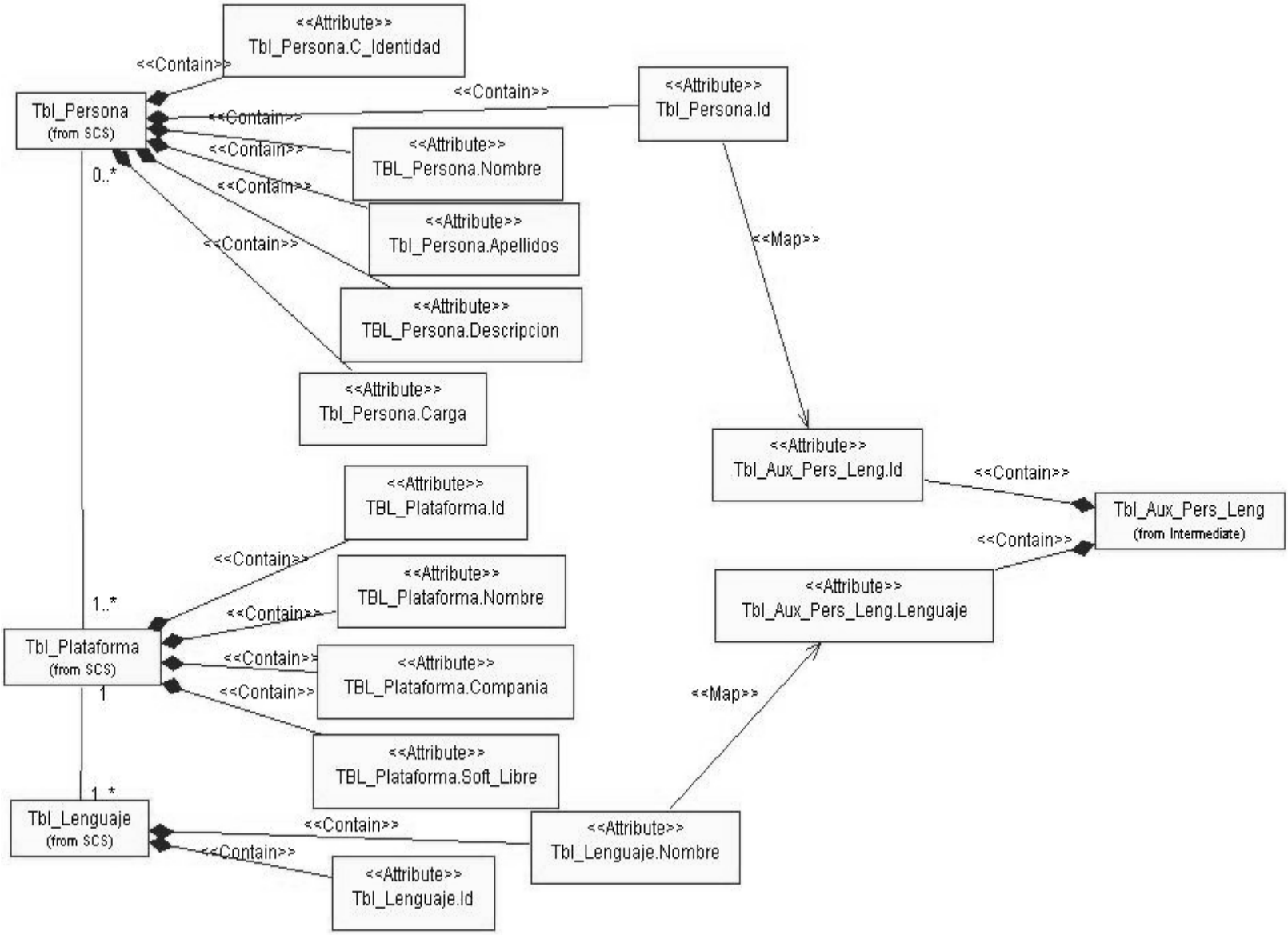
Anexo 3.15 Esquema conceptual del DWH. Paquete hecho Pack Fact_Historial_Persona (nivel 3)



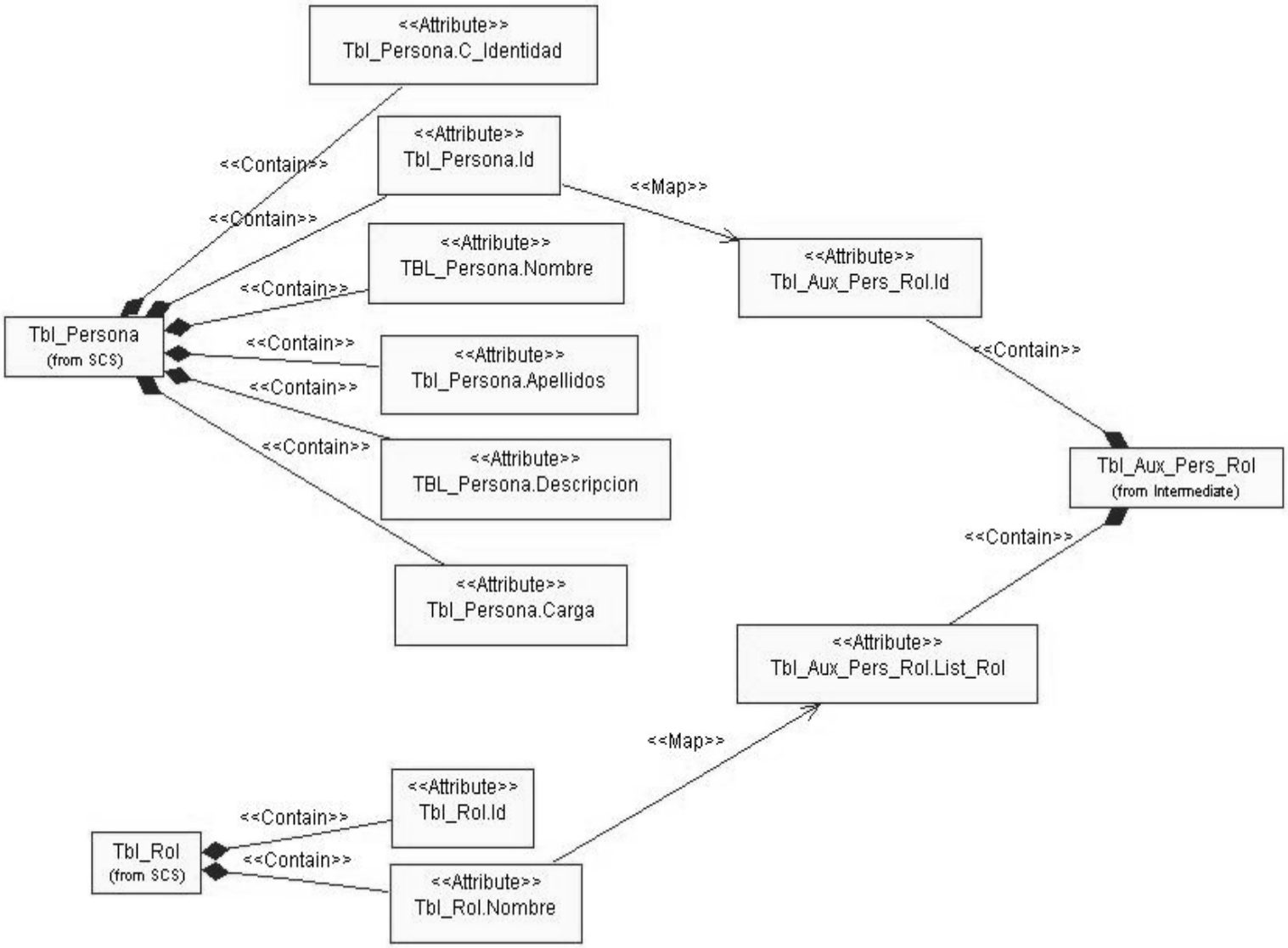
Anexo 3.16 Vista global nivel 2



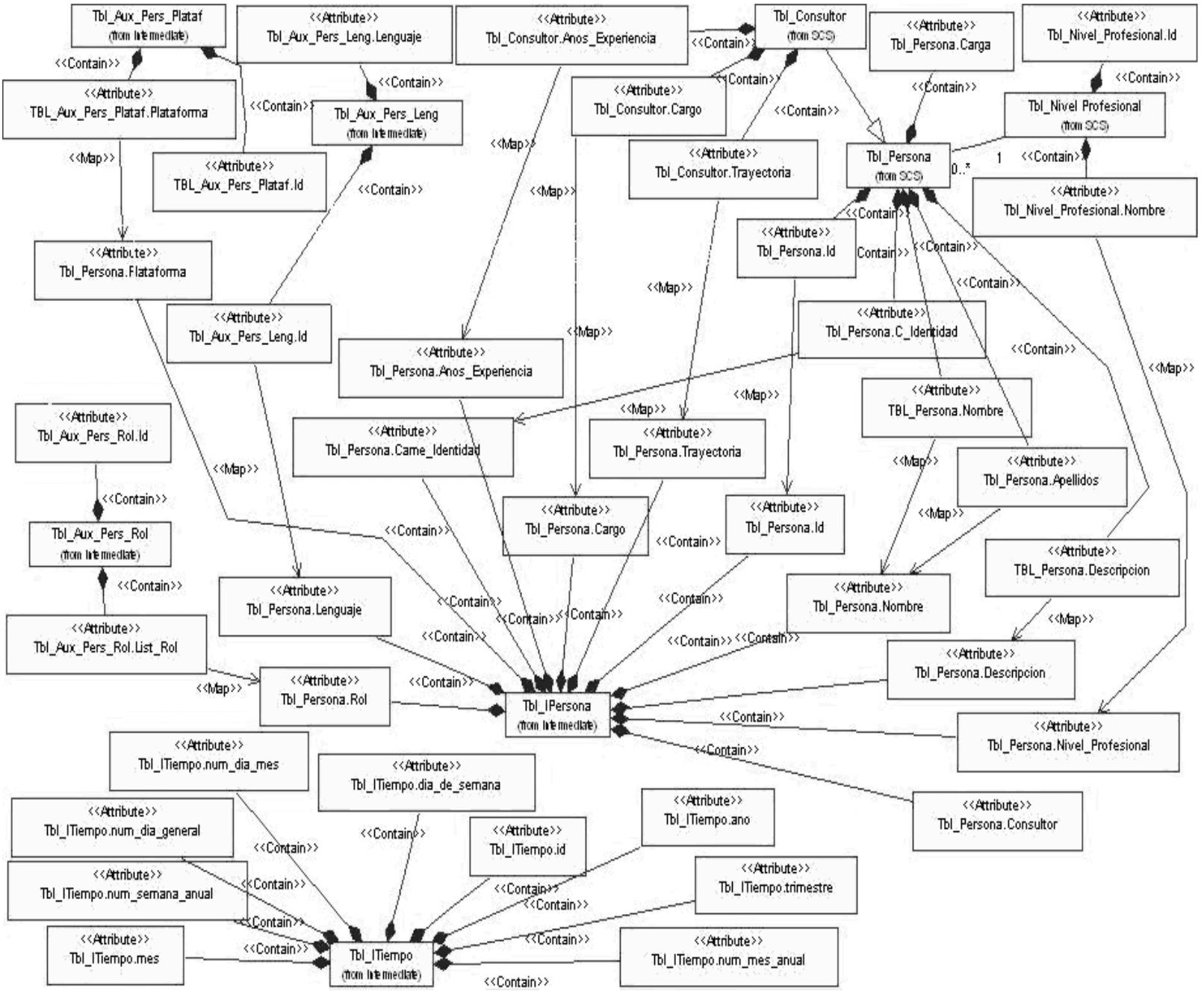
Anexo 3.17 Mapeo1. Paso2 (nivel3)



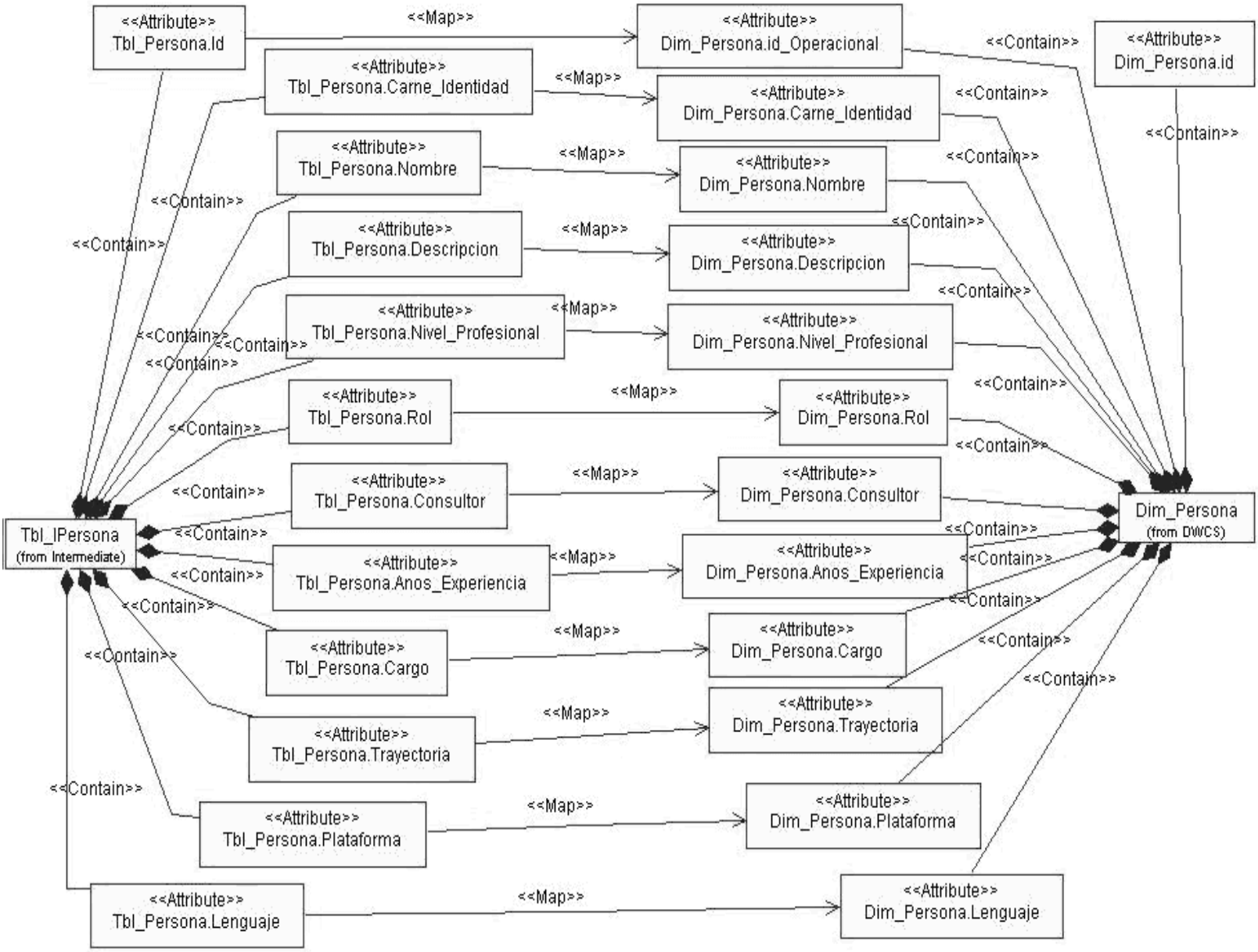
Anexo 3.18 Mapeo1. Paso3 (nivel3)



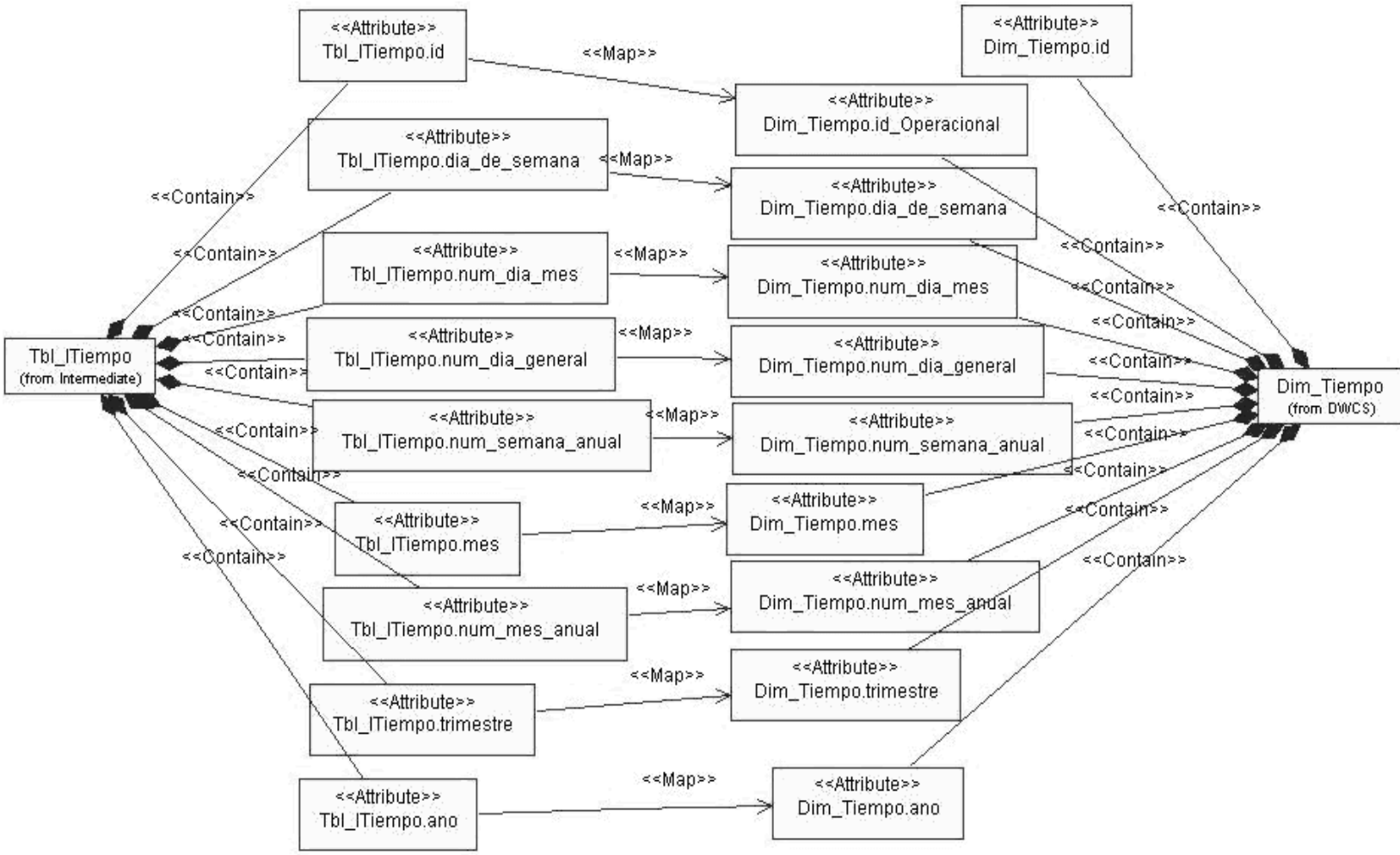
Anexo 3.19 Mapeo1. Paso4 (nivel3)



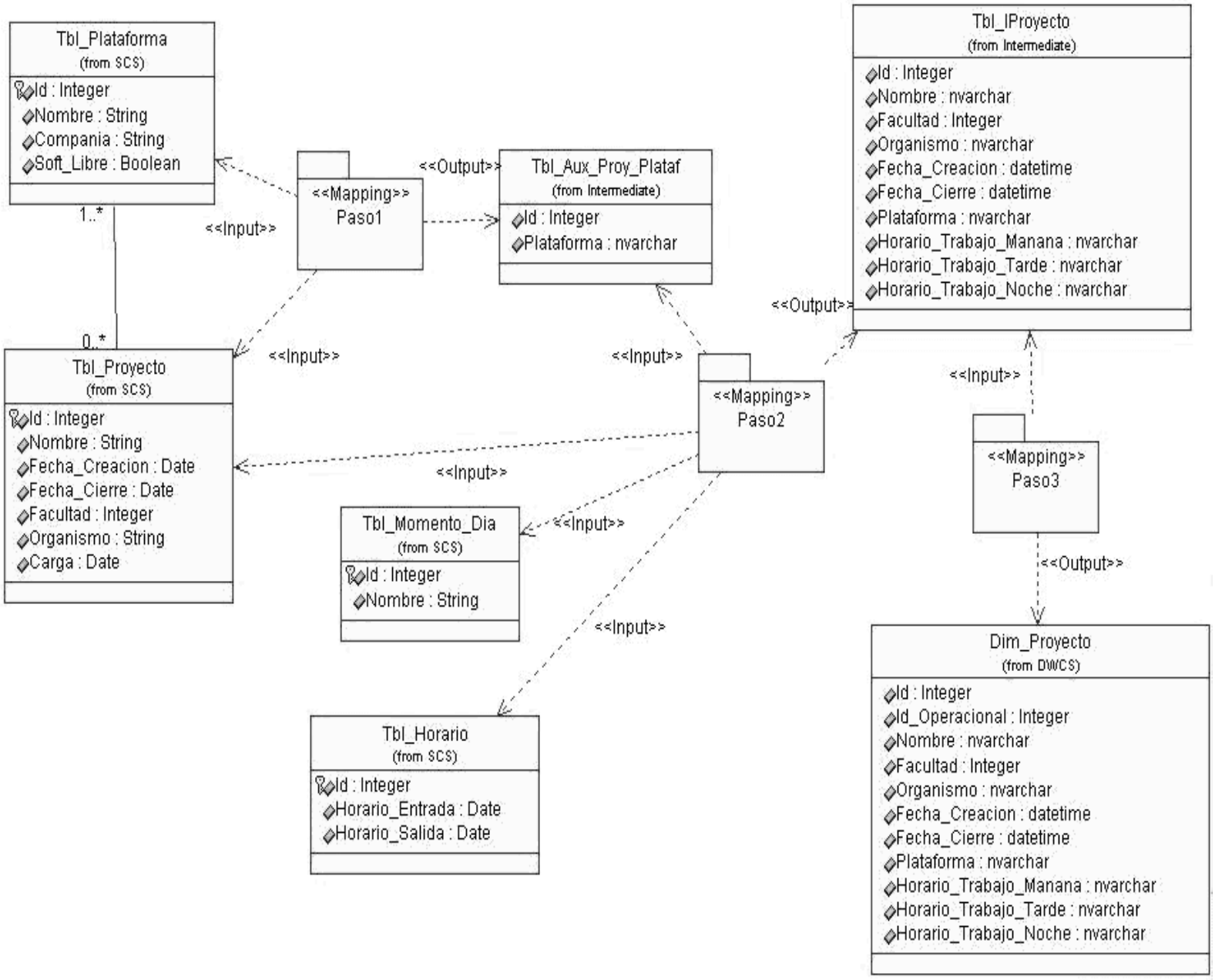
Anexo 3.20 Mapeo1. Paso5 (nivel3)



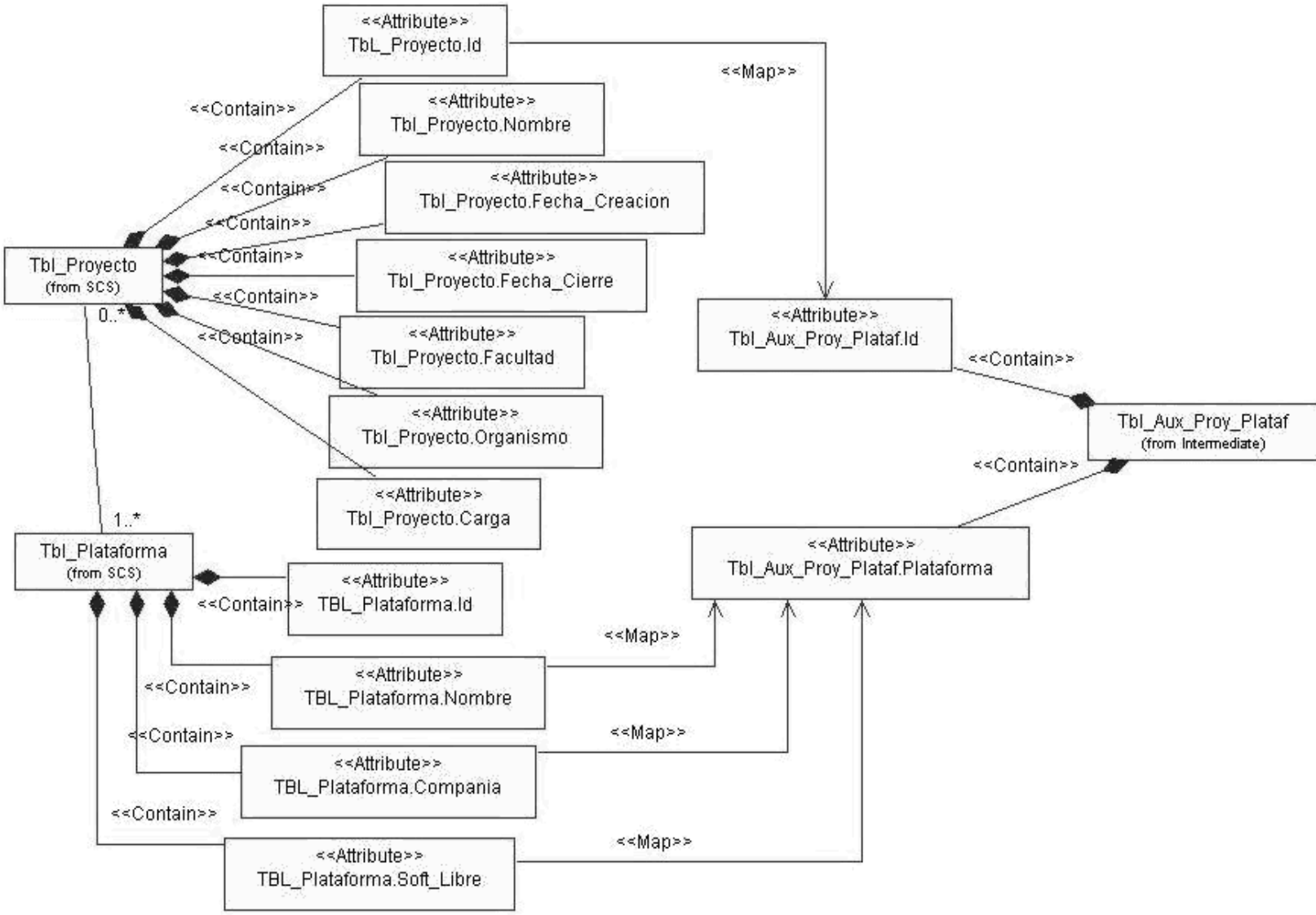
Anexo 3.21 Mapeo1. Paso6 (nivel3)



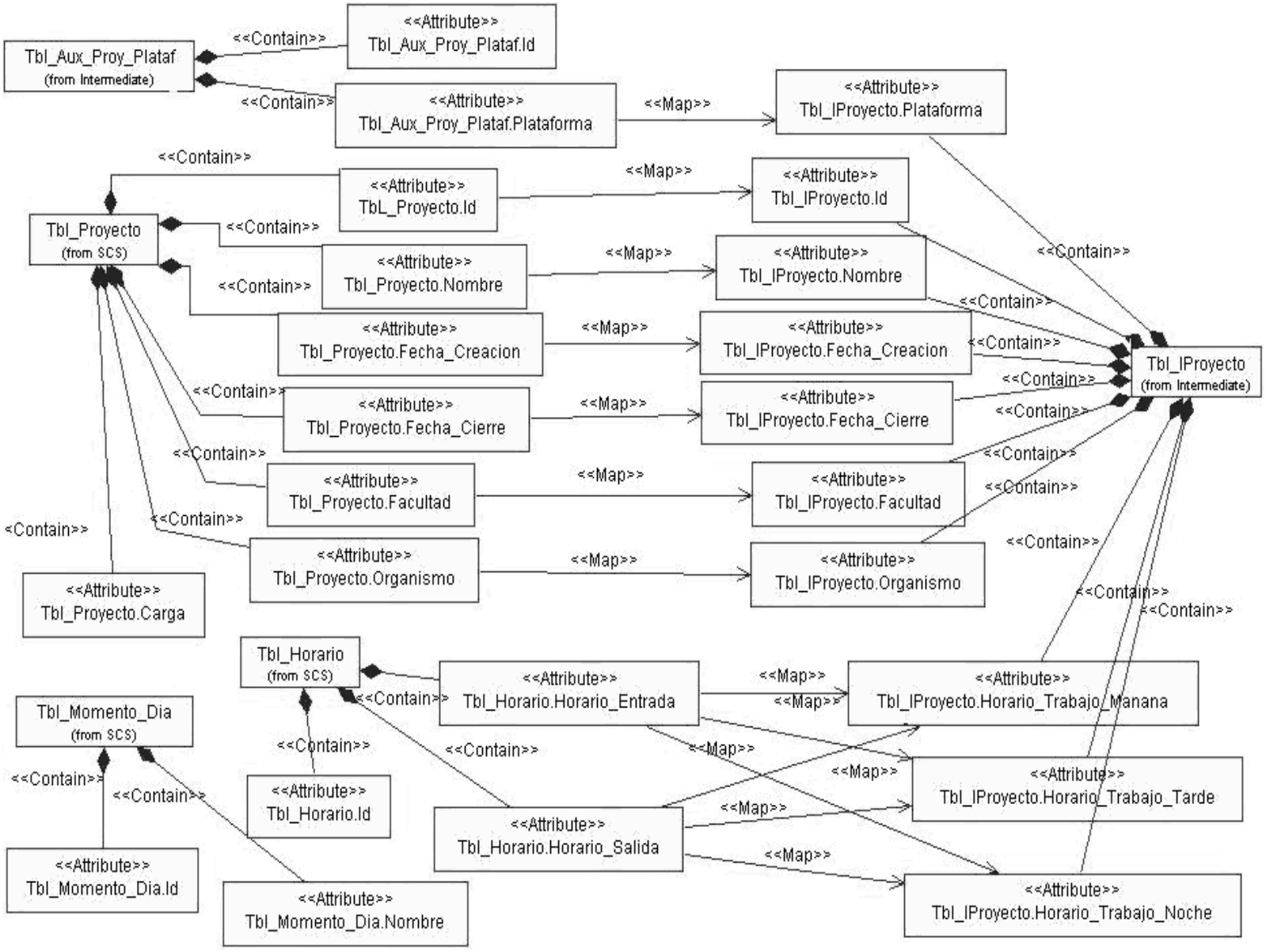
Anexo 3.22 Mapeo2 (nivel2)



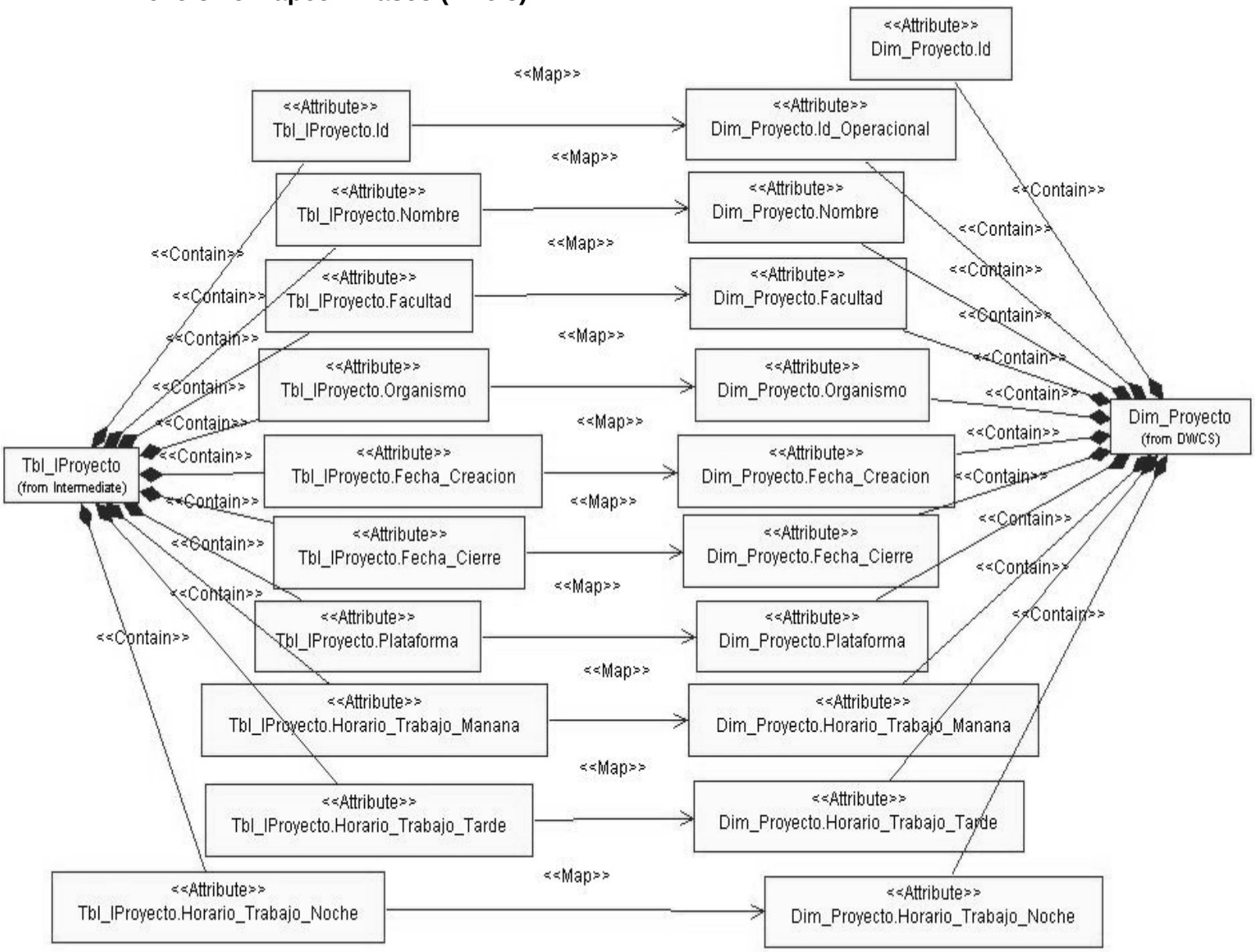
Anexo 3.23 Mapeo2. Paso1 (nivel3)



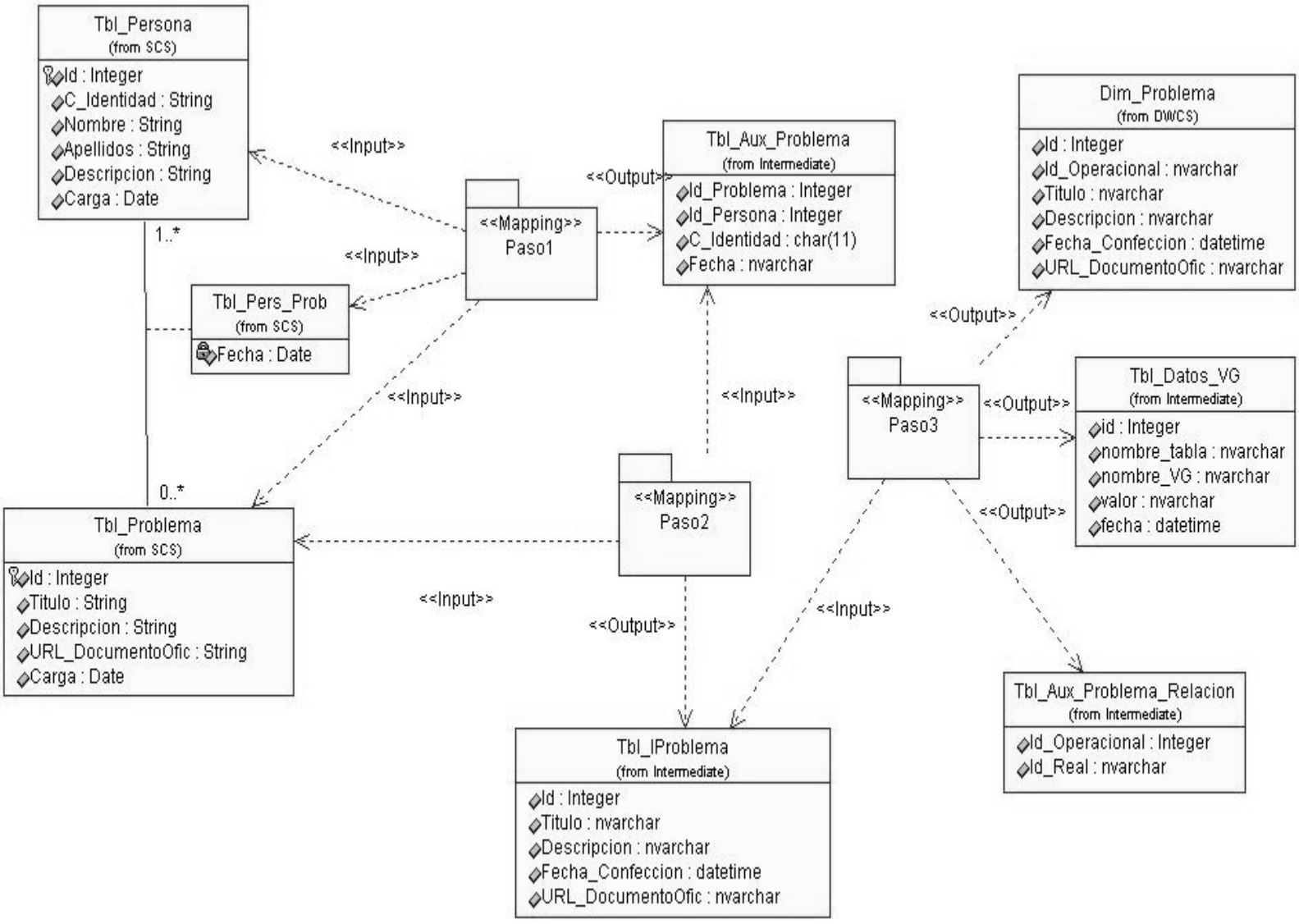
Anexo 3.24 Mapeo2. Paso2 (nivel3)



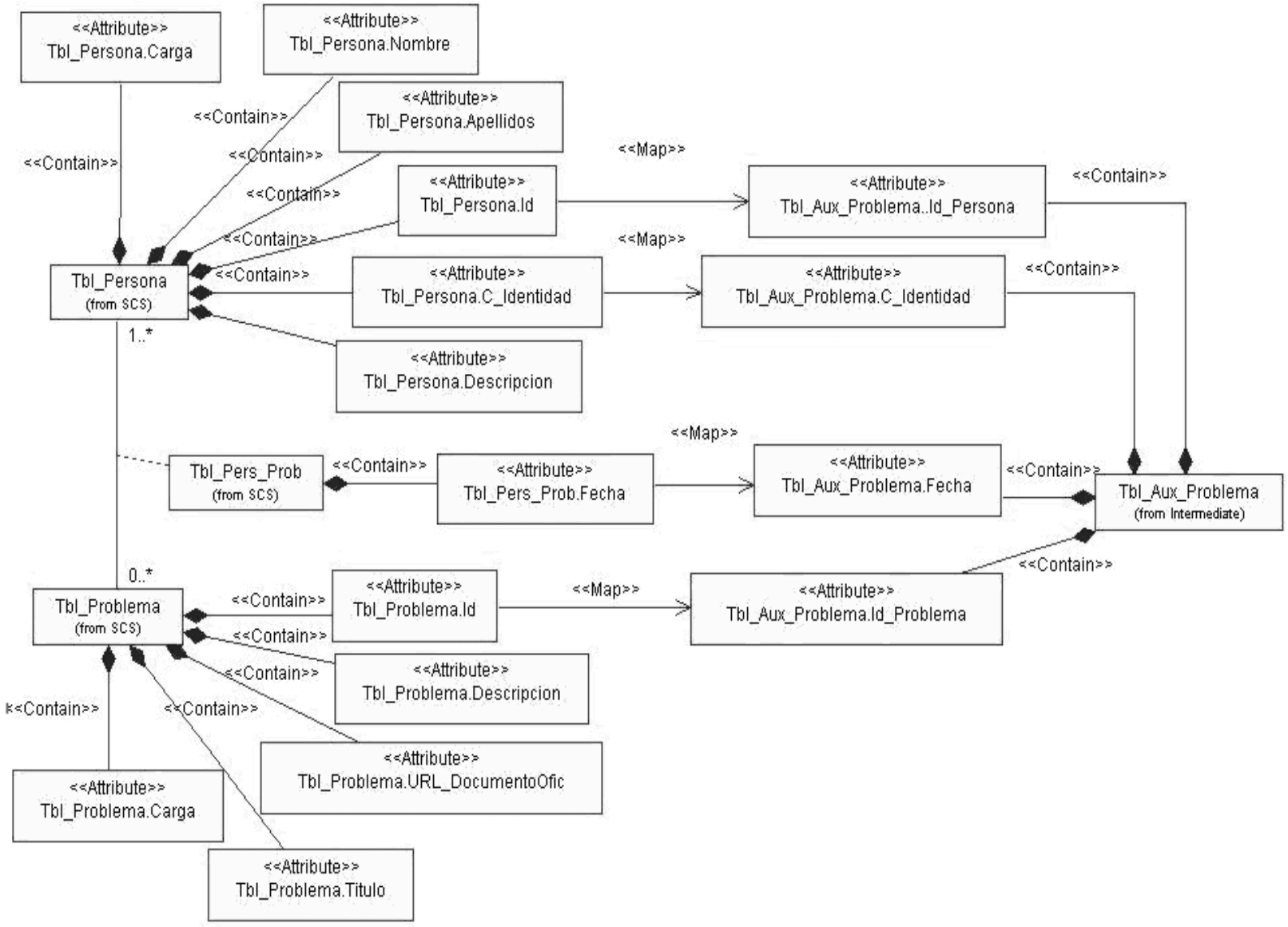
Anexo 3.25 Mapeo2. Paso3 (nivel3)



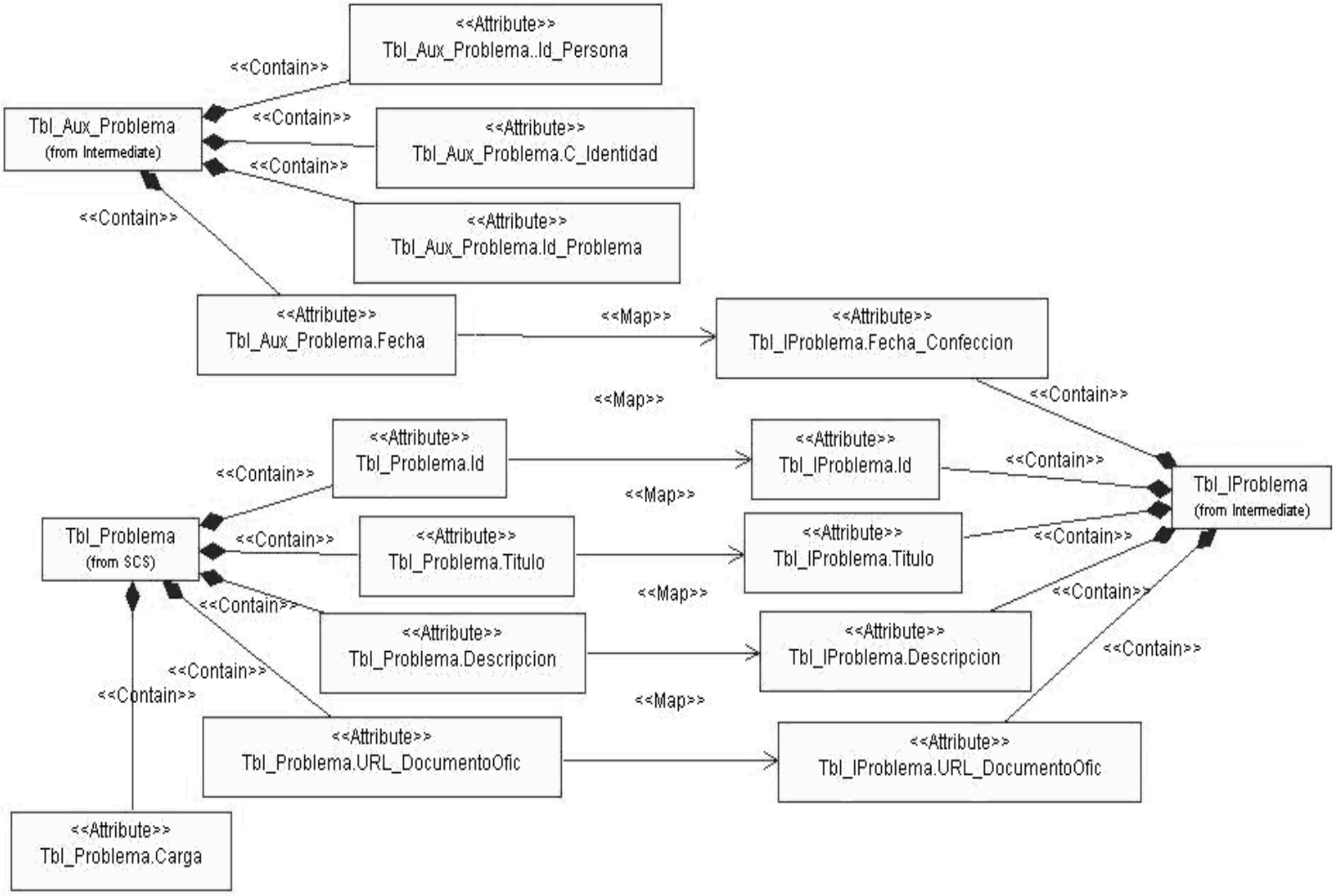
Anexo 3.26 Mapeo3 (nivel2)



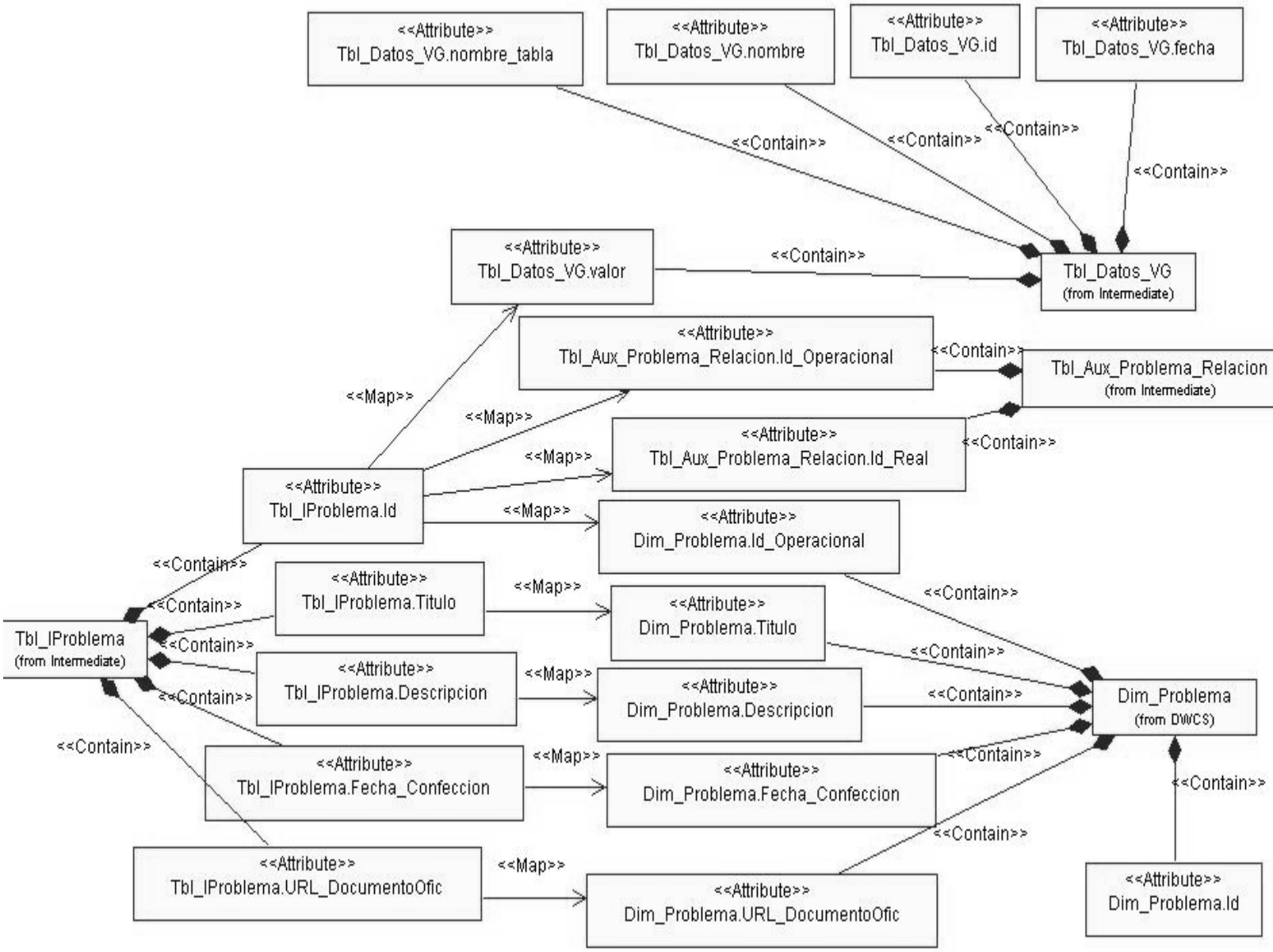
Anexo 3.27 Mapeo3. Paso1 (nivel3)



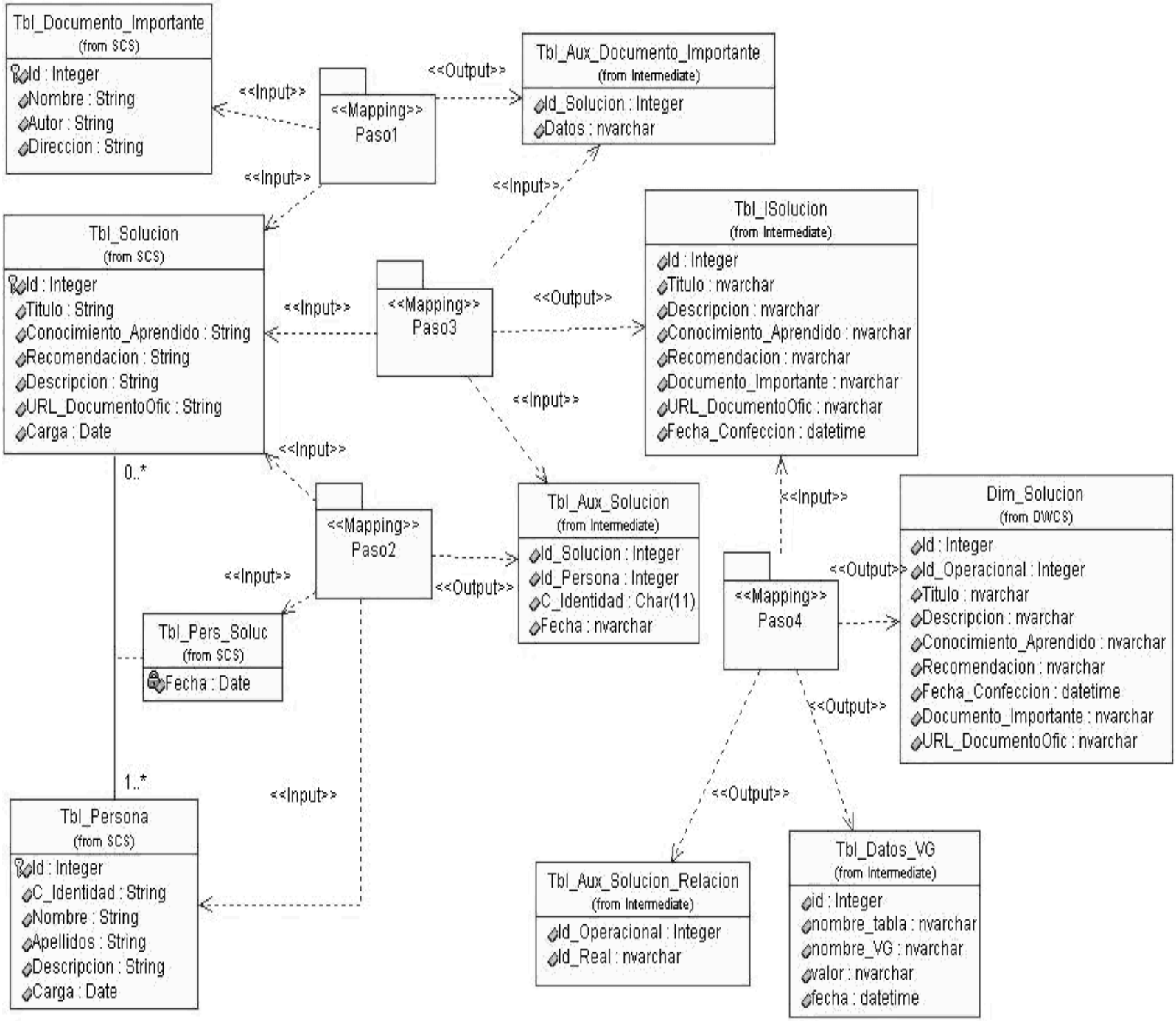
Anexo 3.28 Mapeo3. Paso2 (nivel3)



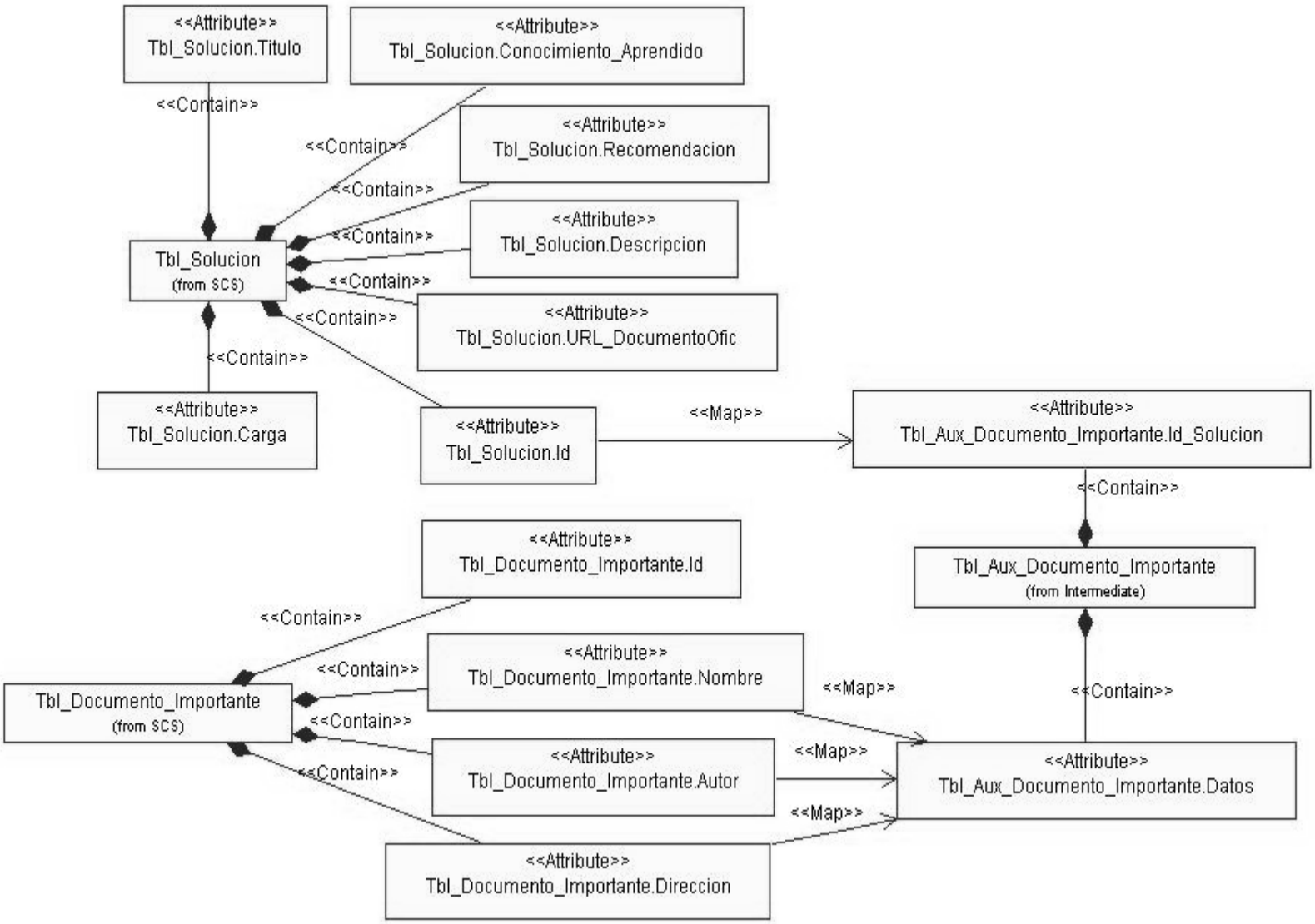
Anexo 3.29 Mapeo3. Paso3 (nivel3)



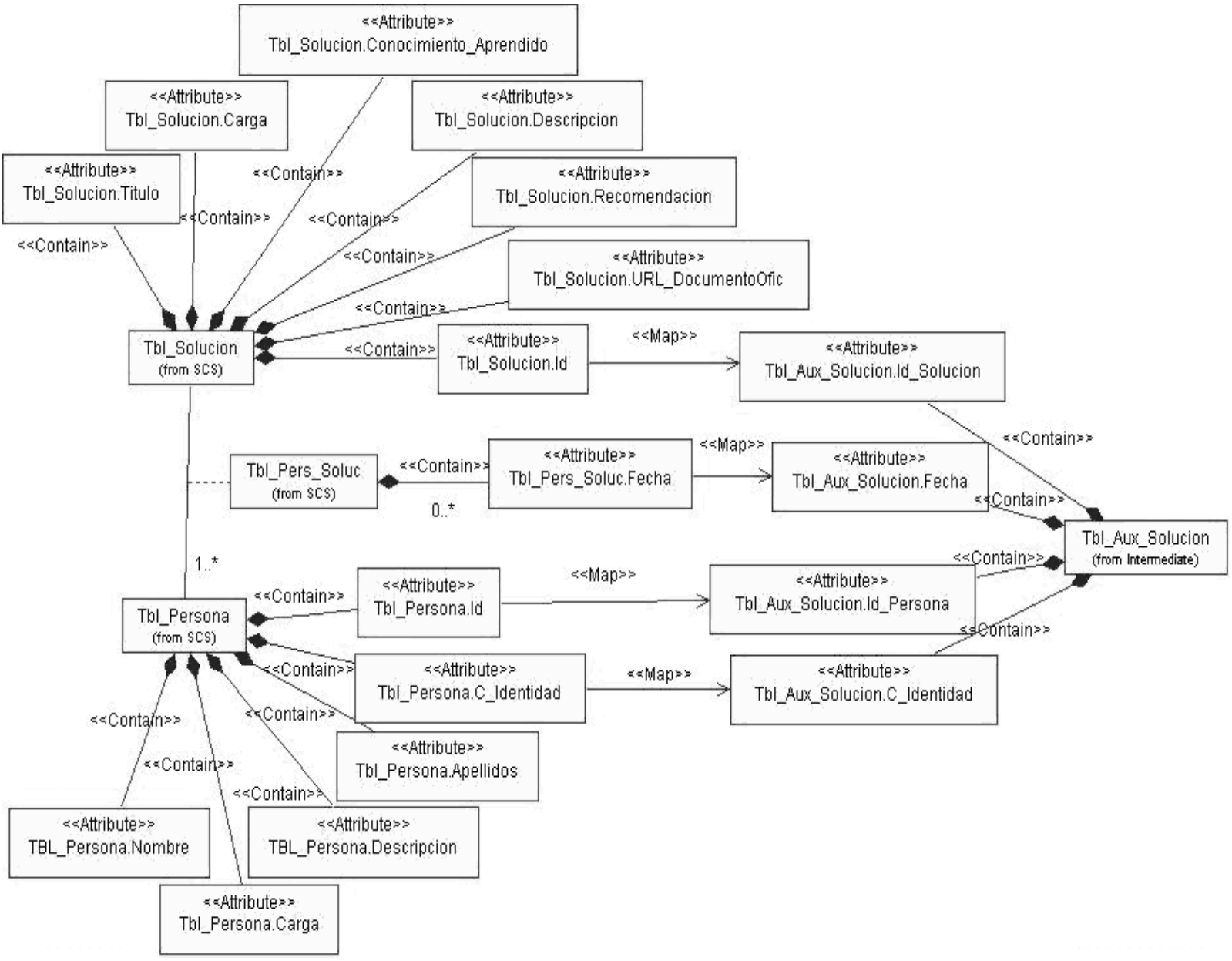
Anexo 3.30 Mapeo 4 (nivel2)



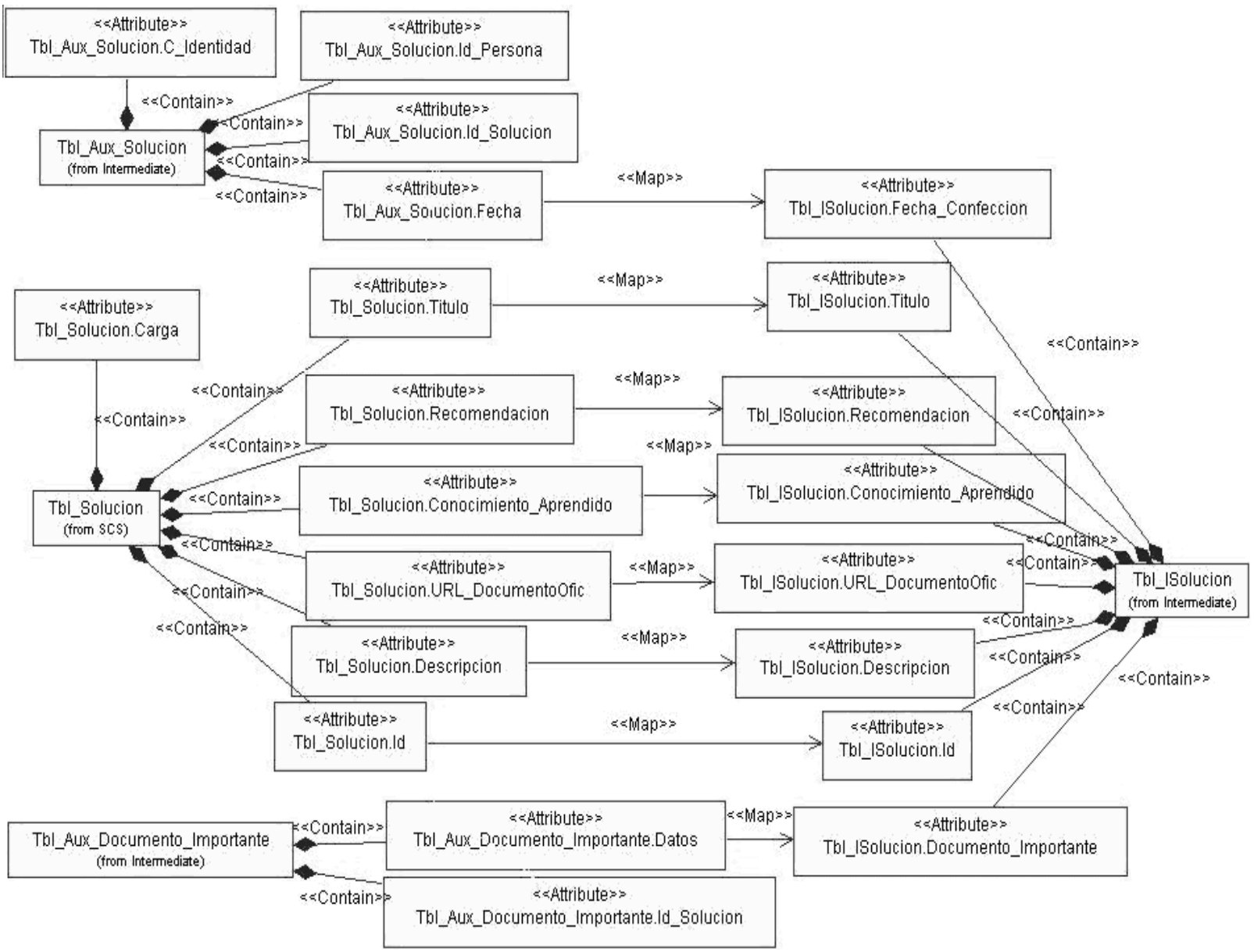
Anexo 3.31 Mapeo4. Paso1 (nivel3)



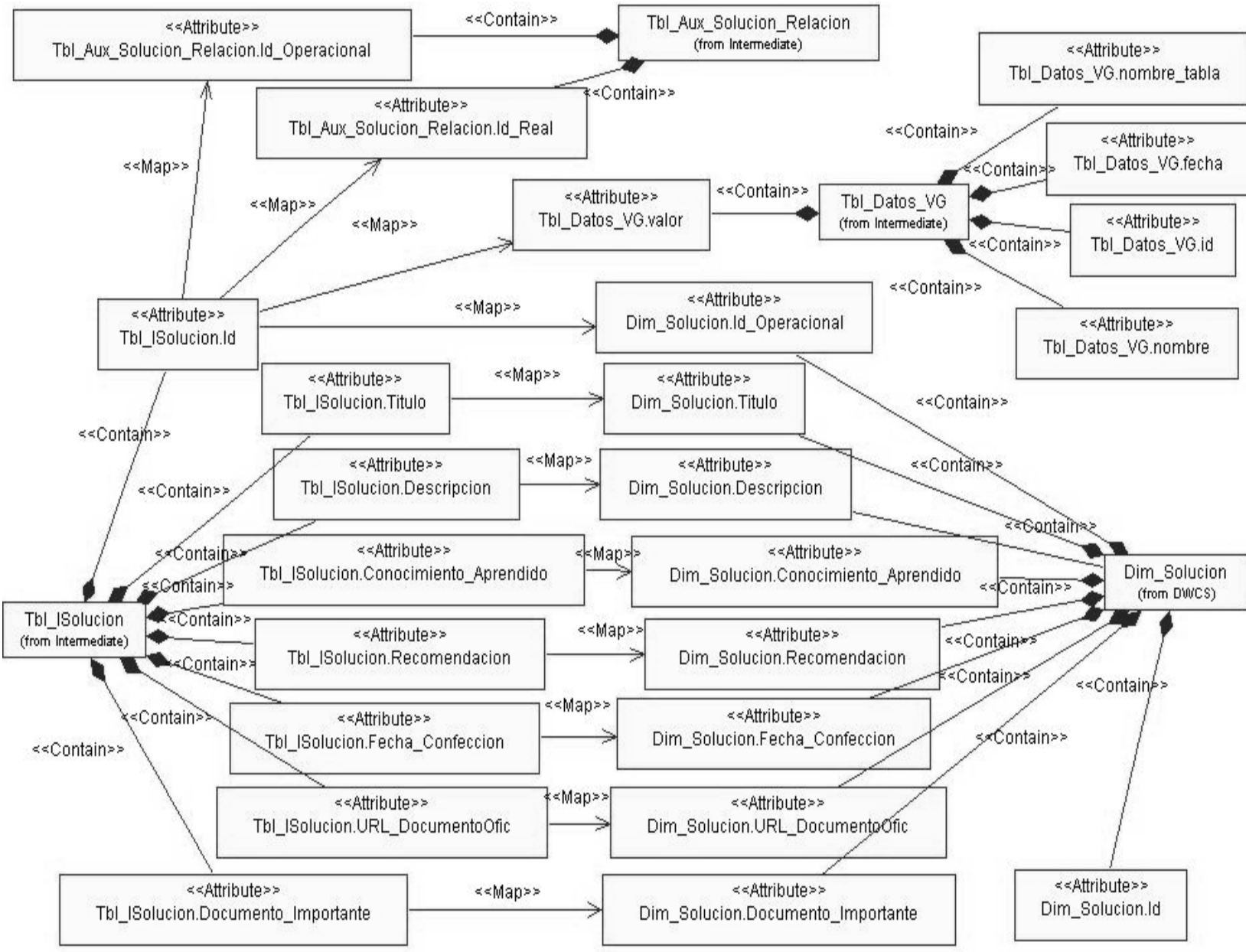
Anexo 3.32 Mapeo4. Paso2 (nivel3)



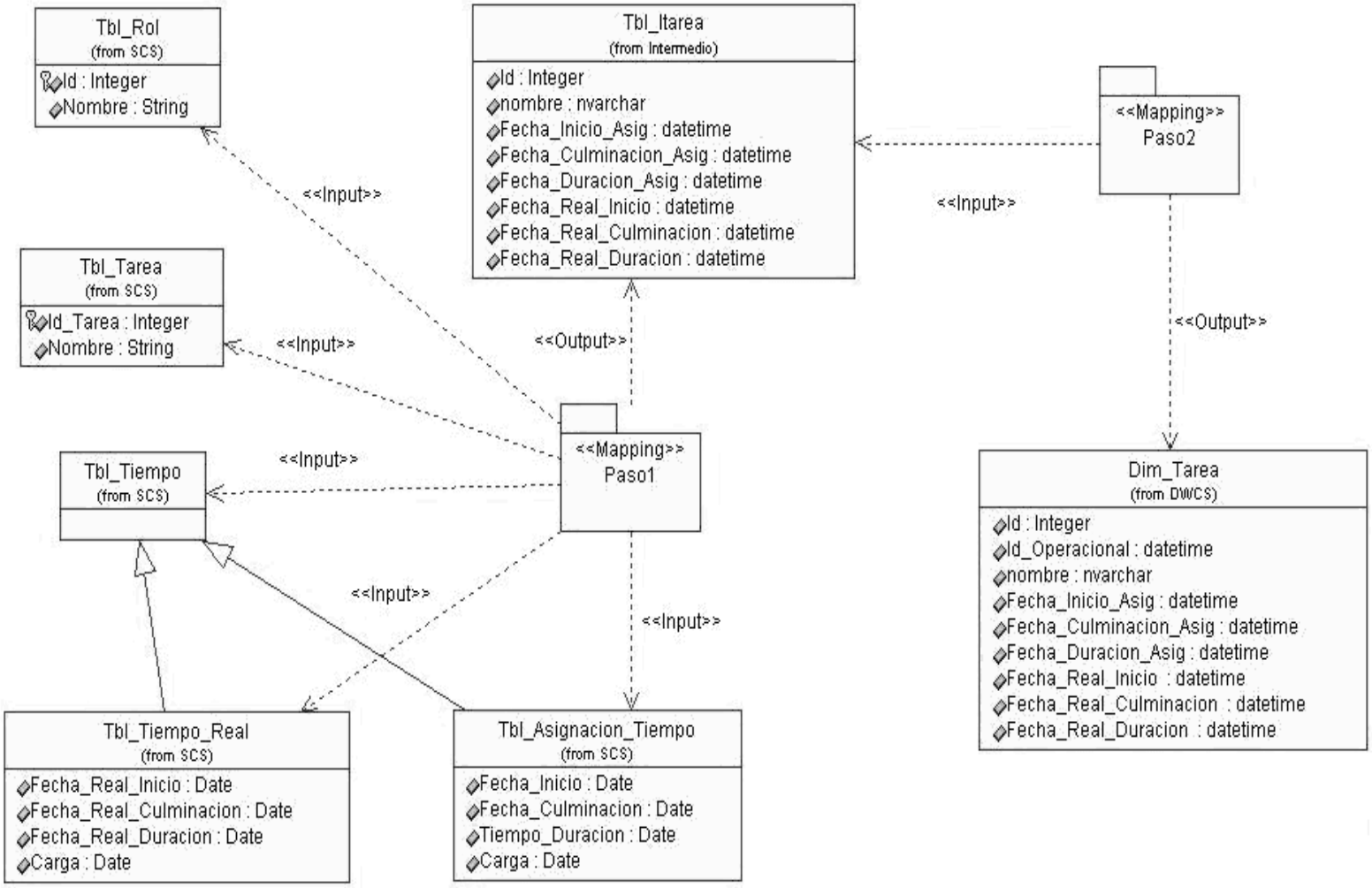
Anexo 3.33 Mapeo4. Paso3 (nivel3)



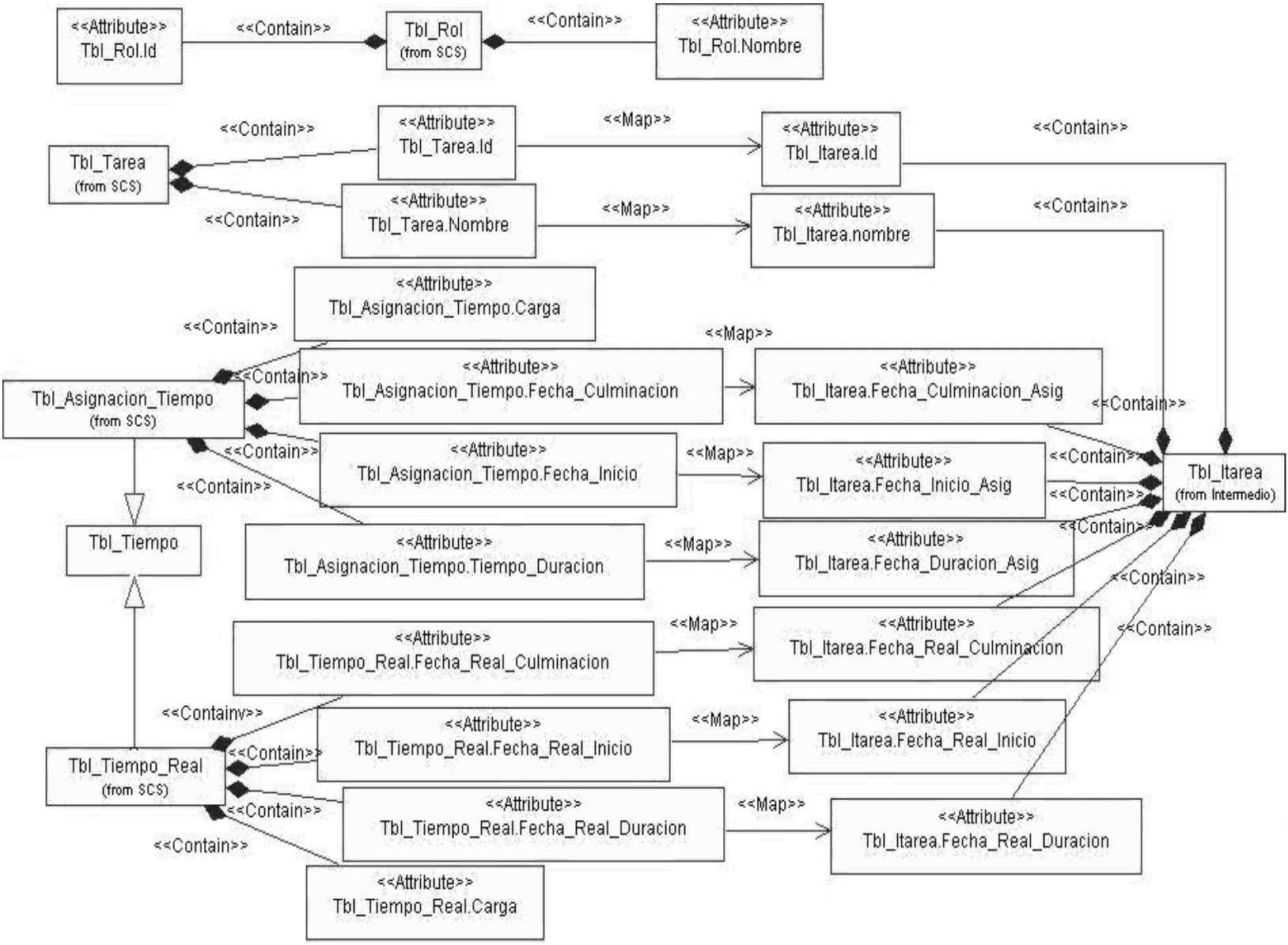
Anexo 3.34 Mapeo4. Paso4 (nivel3)



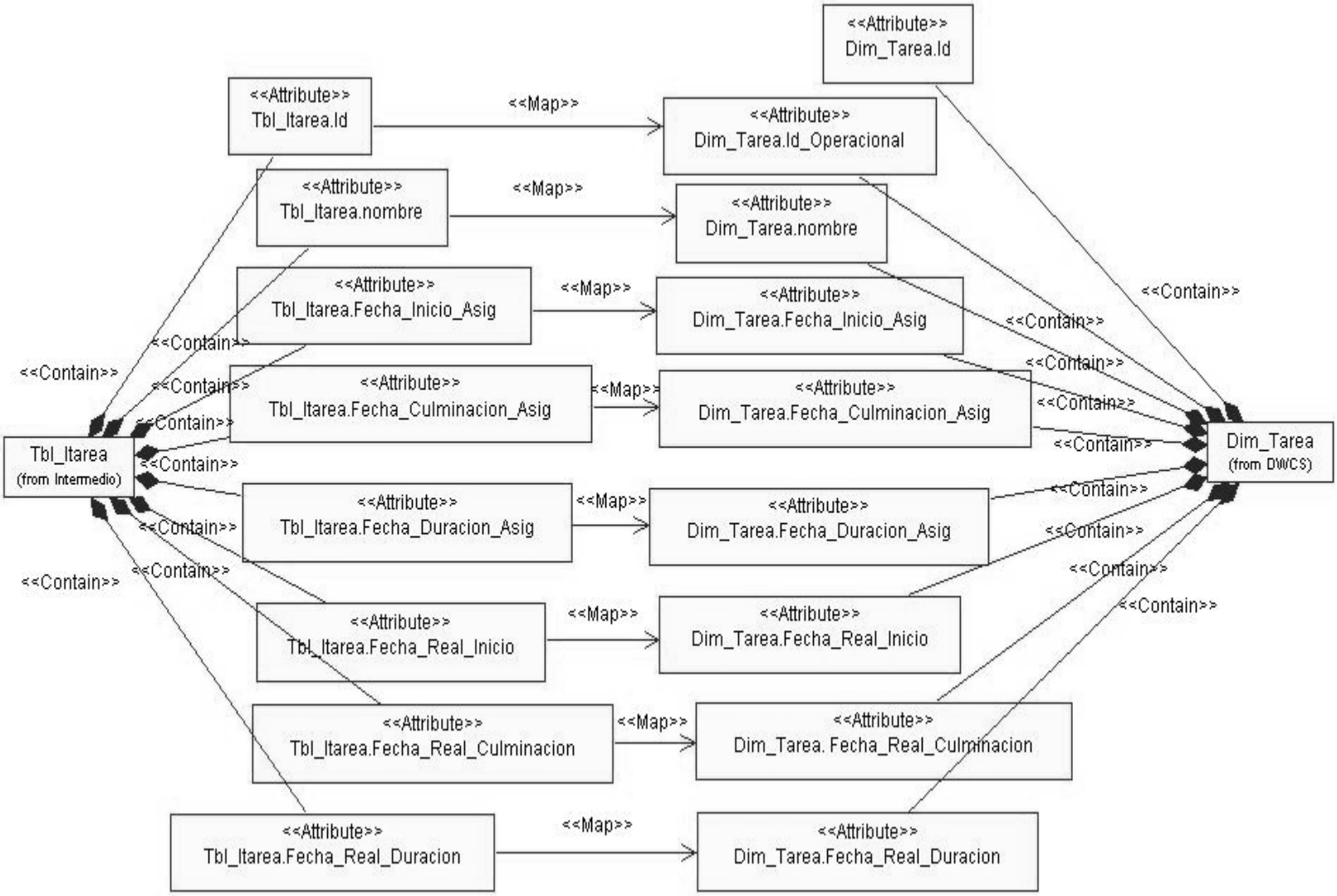
Anexo 3.35 Mapeo5 (nivel2)



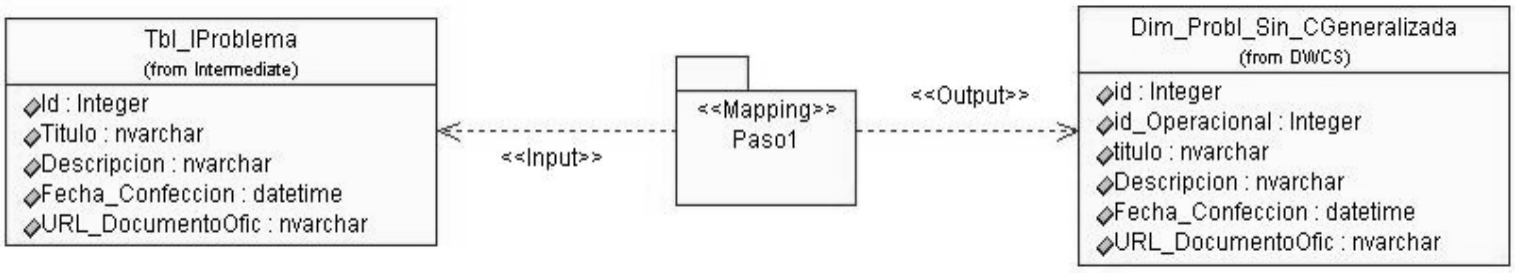
Anexo 3.36 Mapeo5. Paso1 (nivel3)



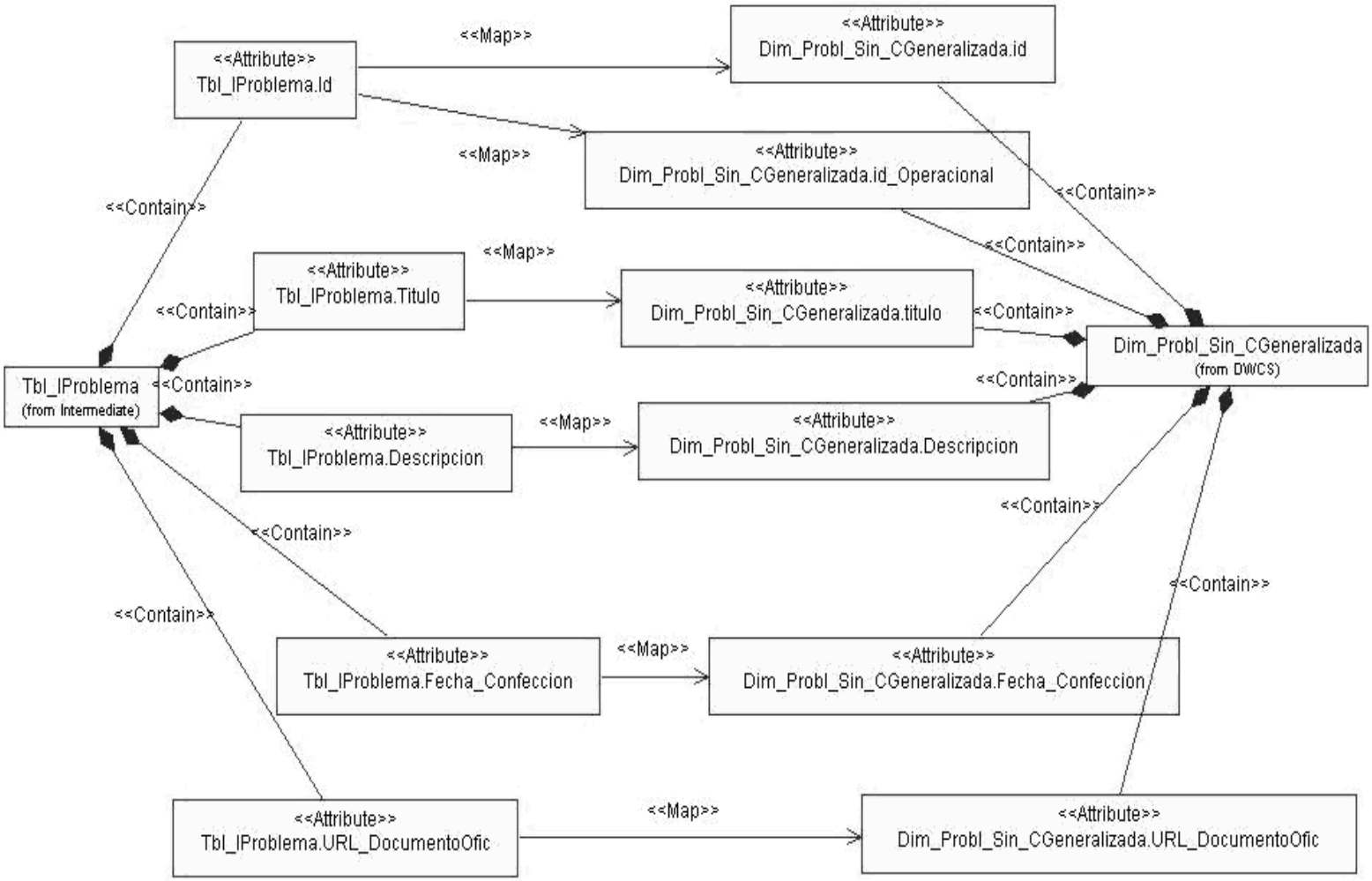
Anexo 3.37 Mapeo5. Paso2 (nivel3)



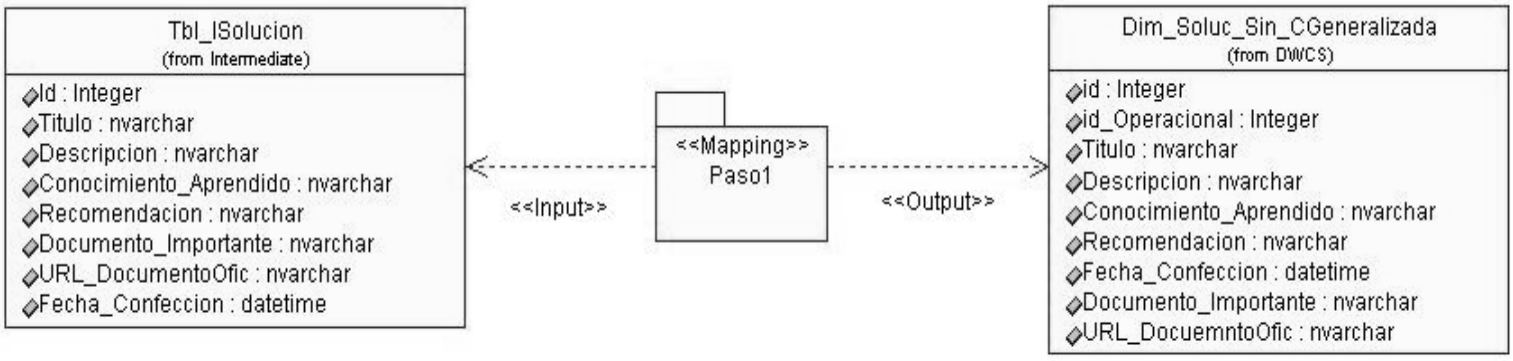
Anexo 3.38 Mapeo6 (nivel2)



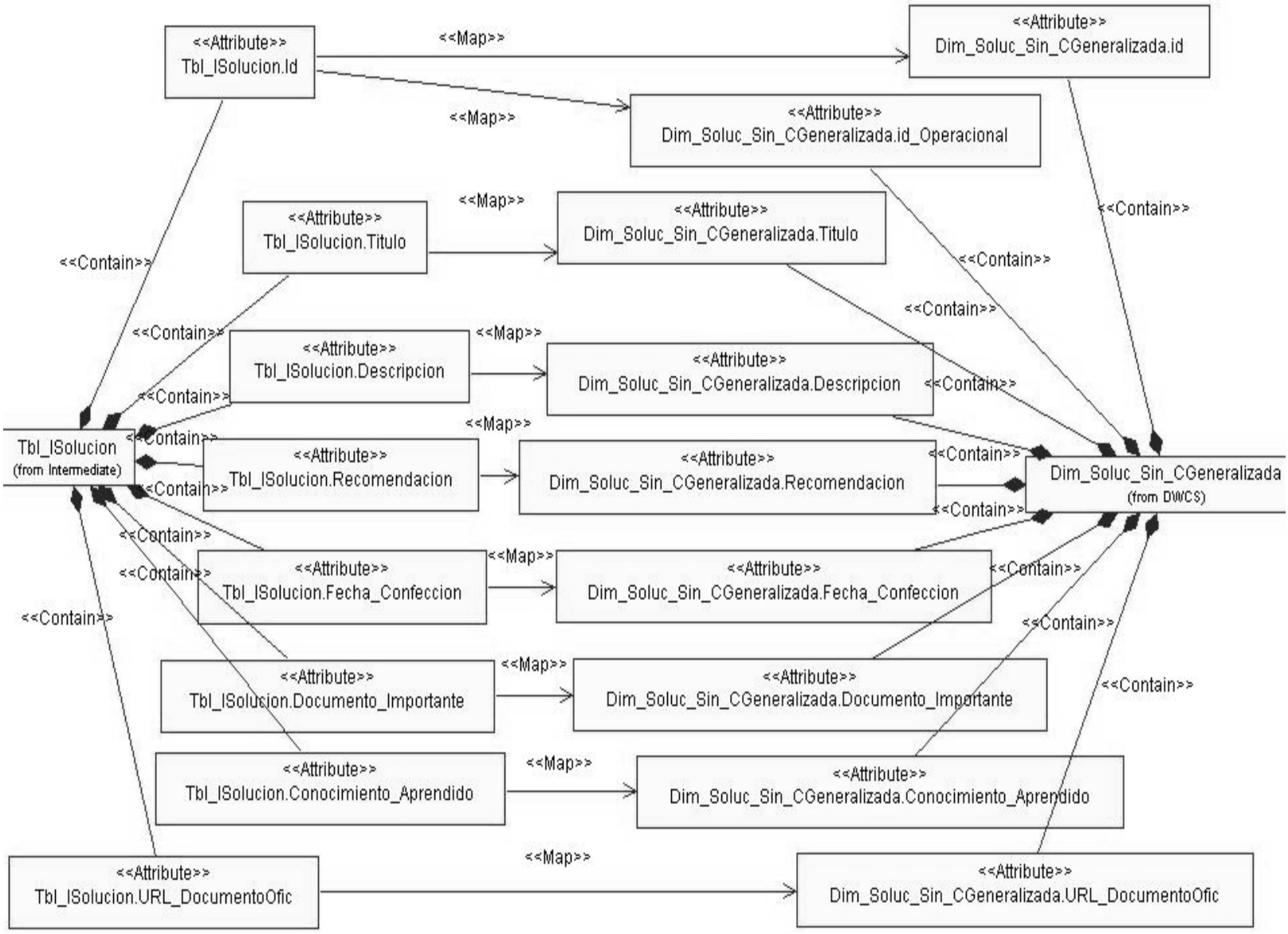
Anexo 3.39 Mapeo6. Paso1 (nivel3)



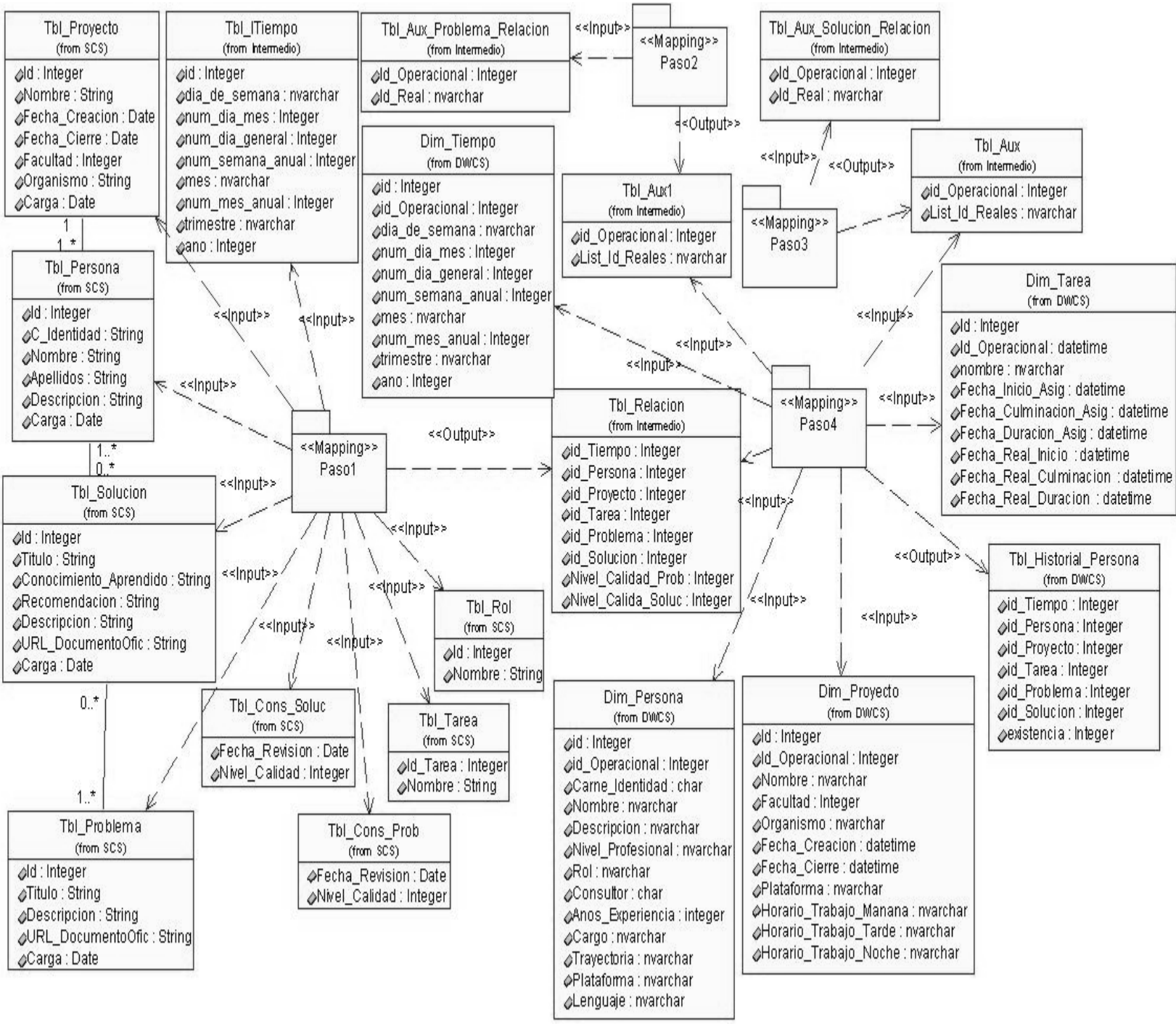
Anexo 3.40 Mapeo7 (nivel2)



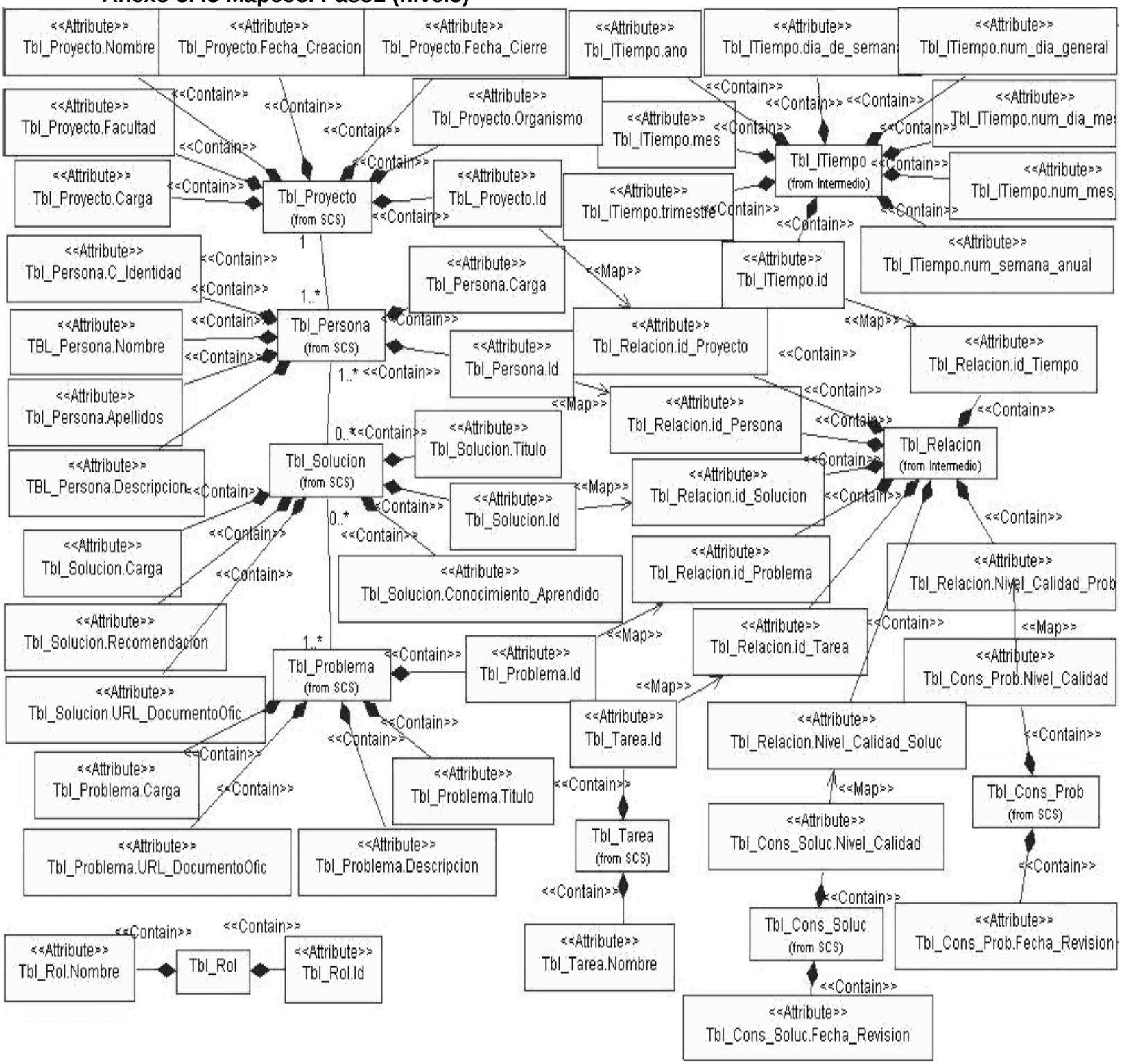
Anexo 3.41 Mapeo7. Paso1 (nivel3)



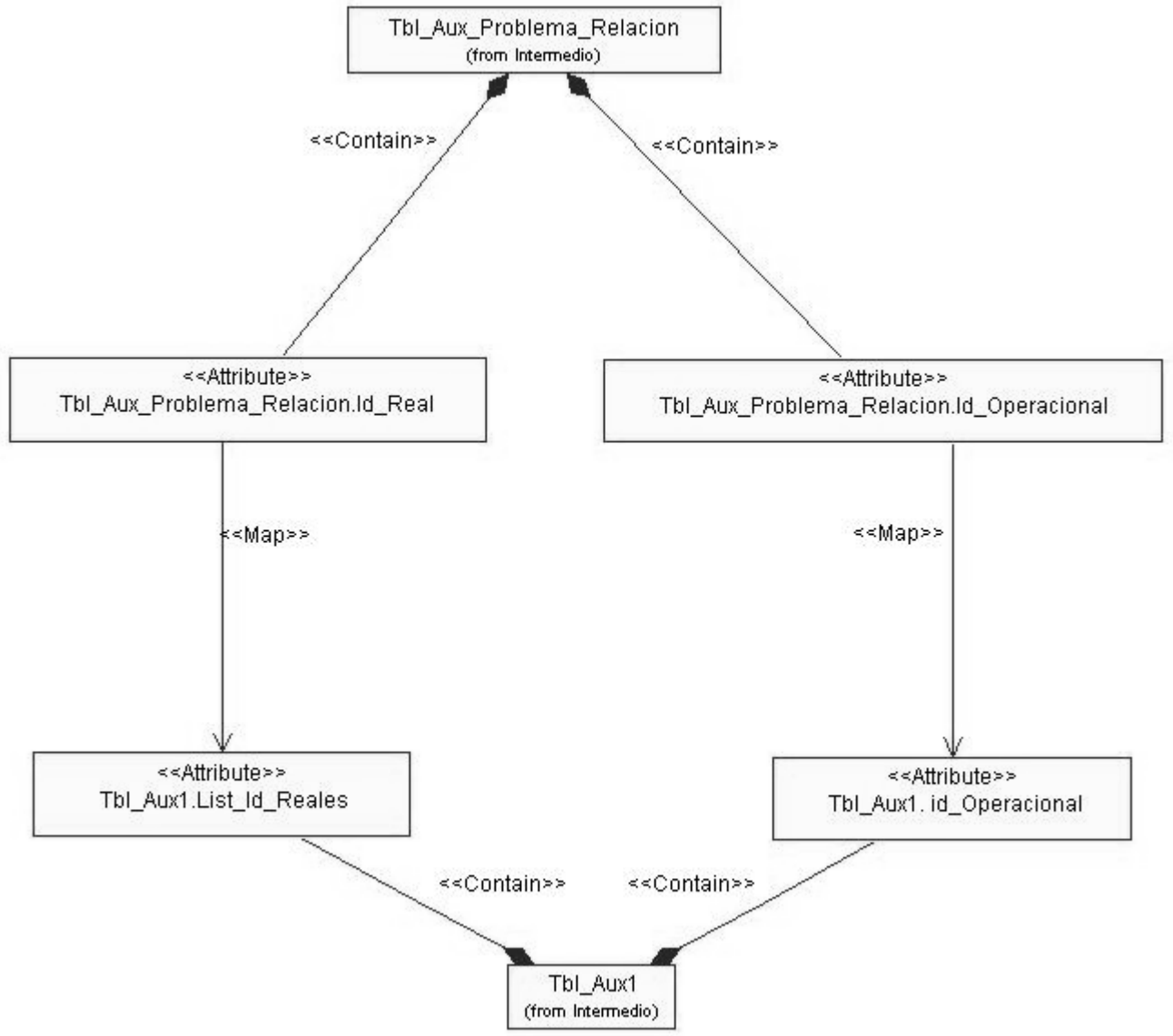
Anexo 3.42 Mapeo8 (nivel2)



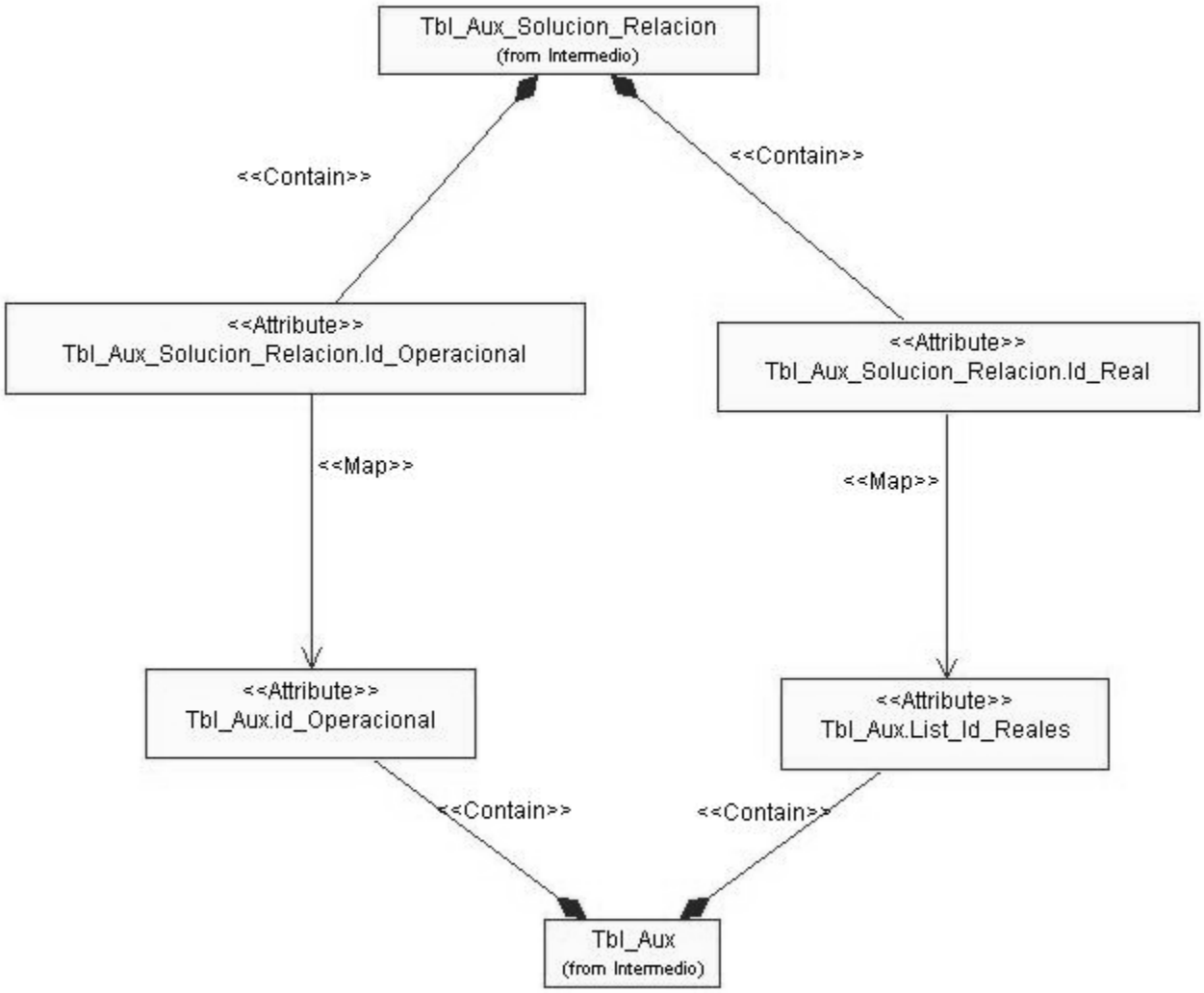
Anexo 3.43 Mapeo8. Paso1 (nivel3)



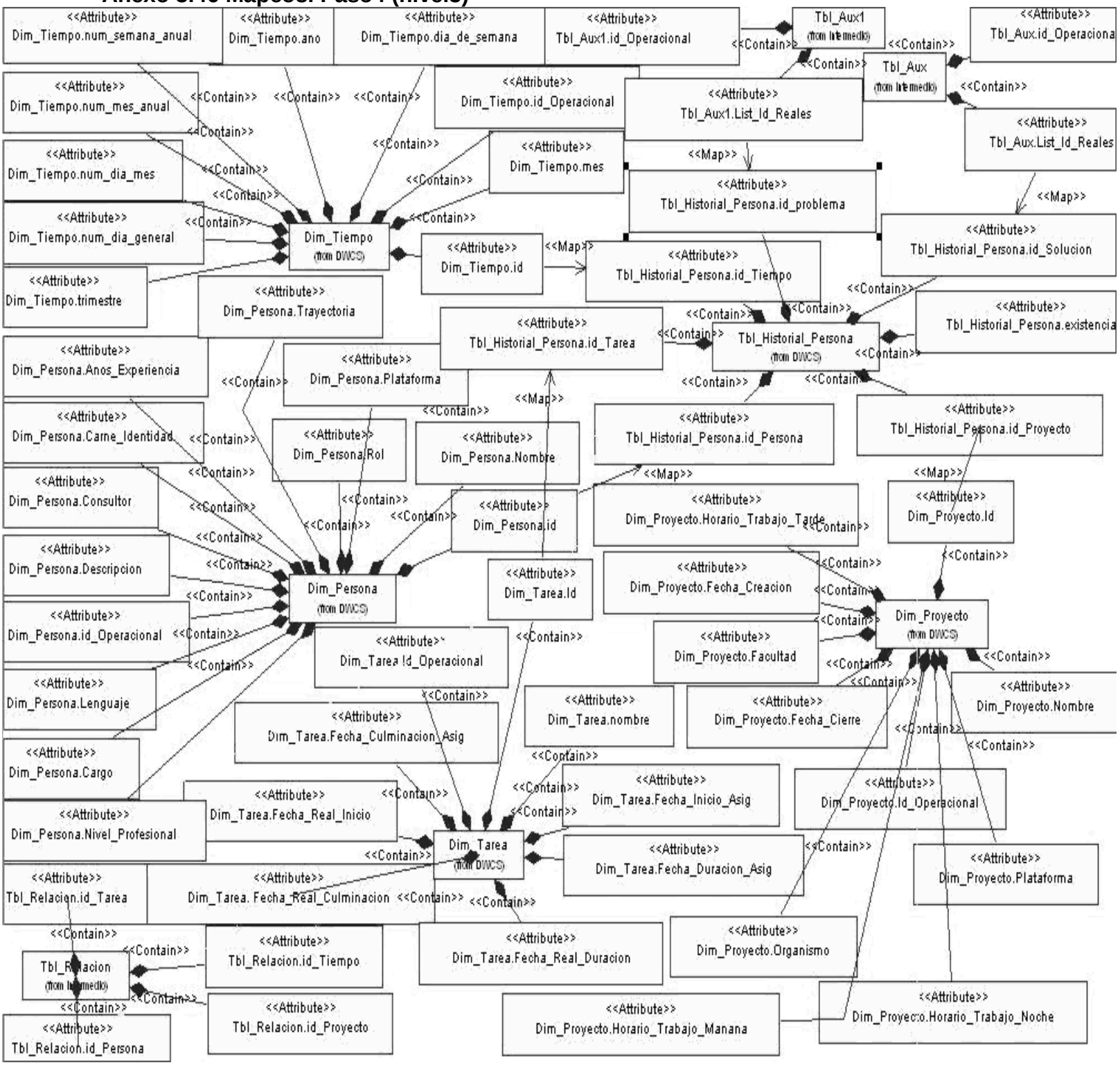
Anexo 3.44 Mapeo8. Paso2 (nivel3)



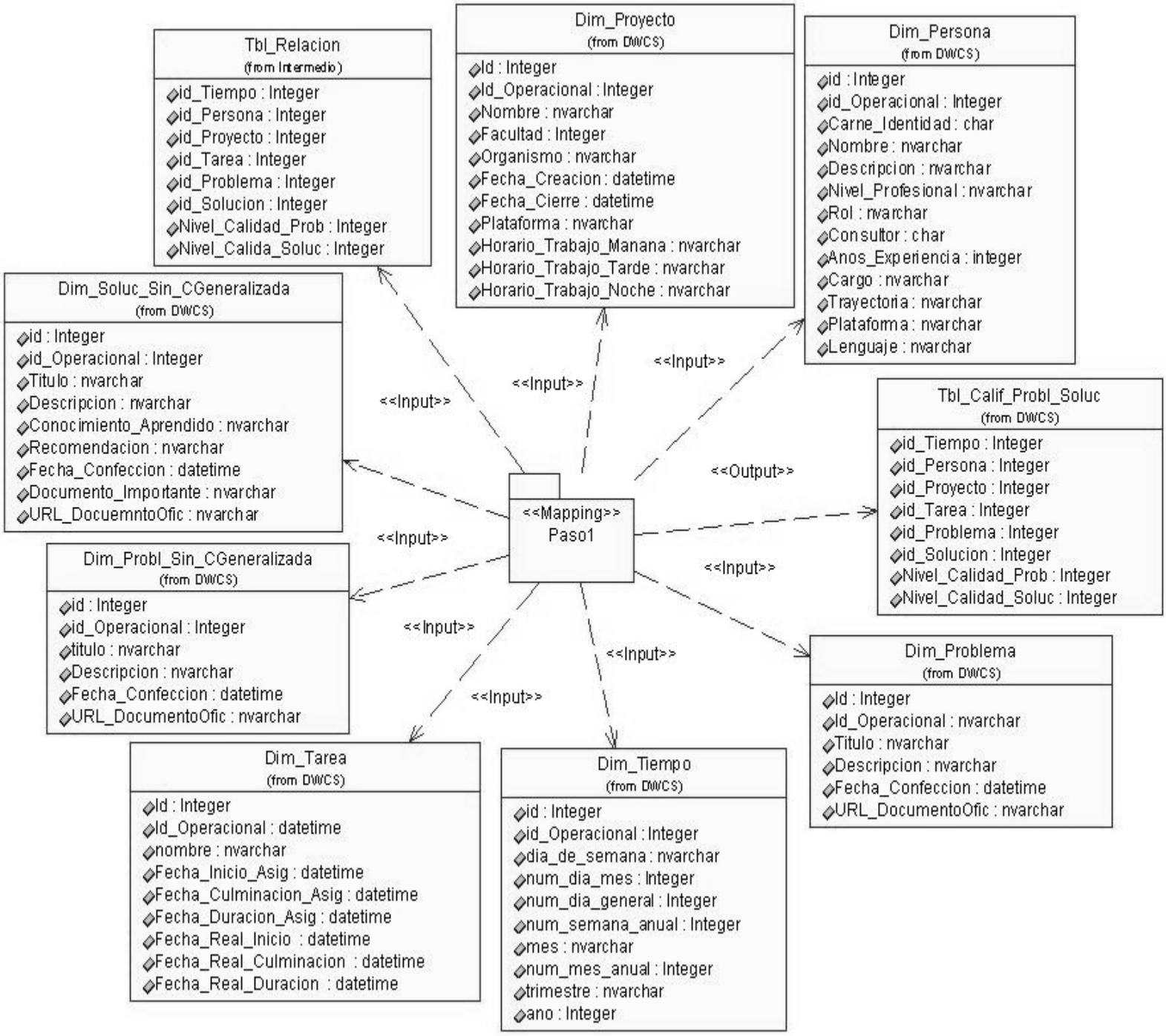
Anexo 3.45 Mapeo8. Paso3 (nivel3)



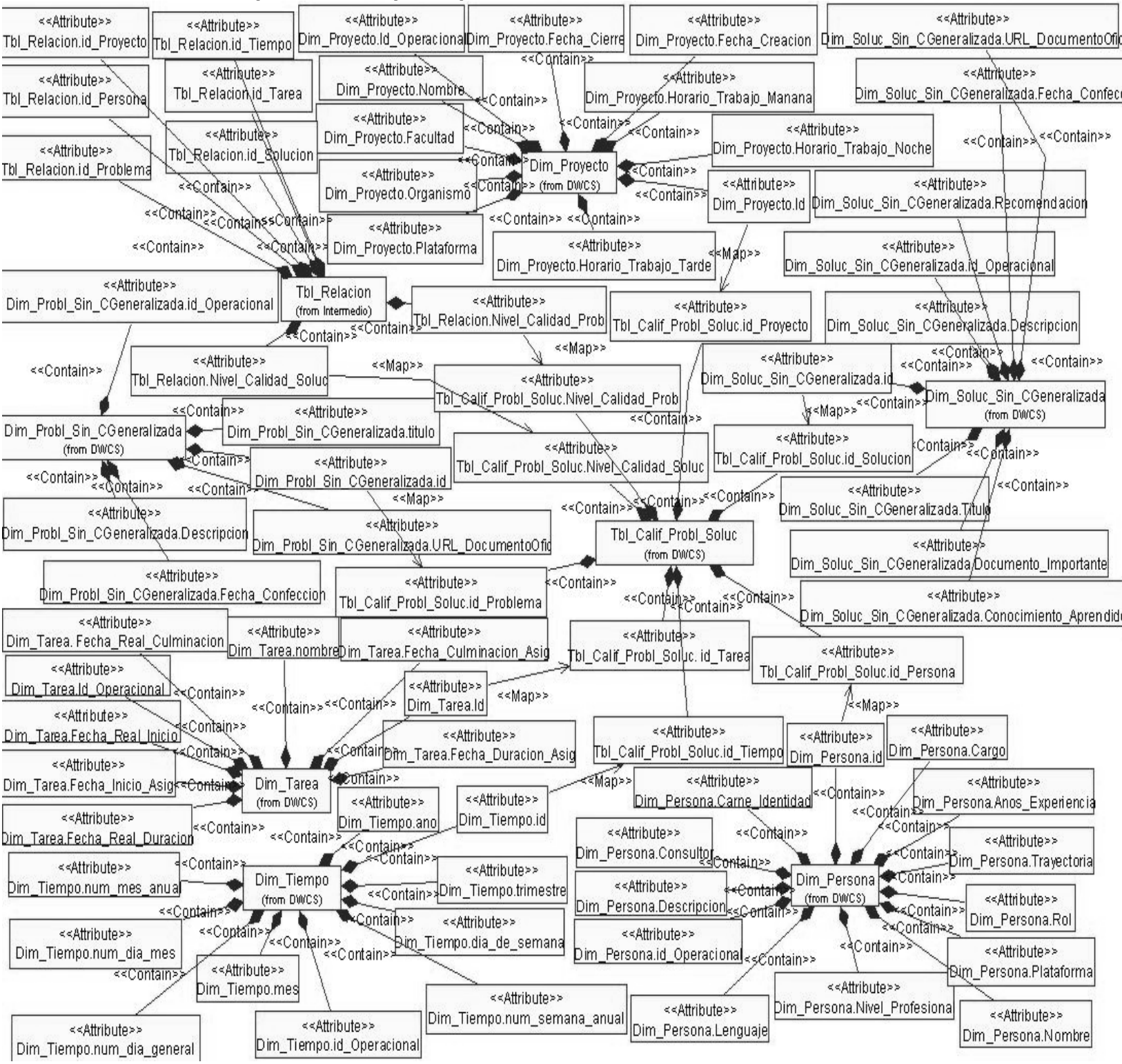
Anexo 3.46 Mapeo8. Paso4 (nivel3)



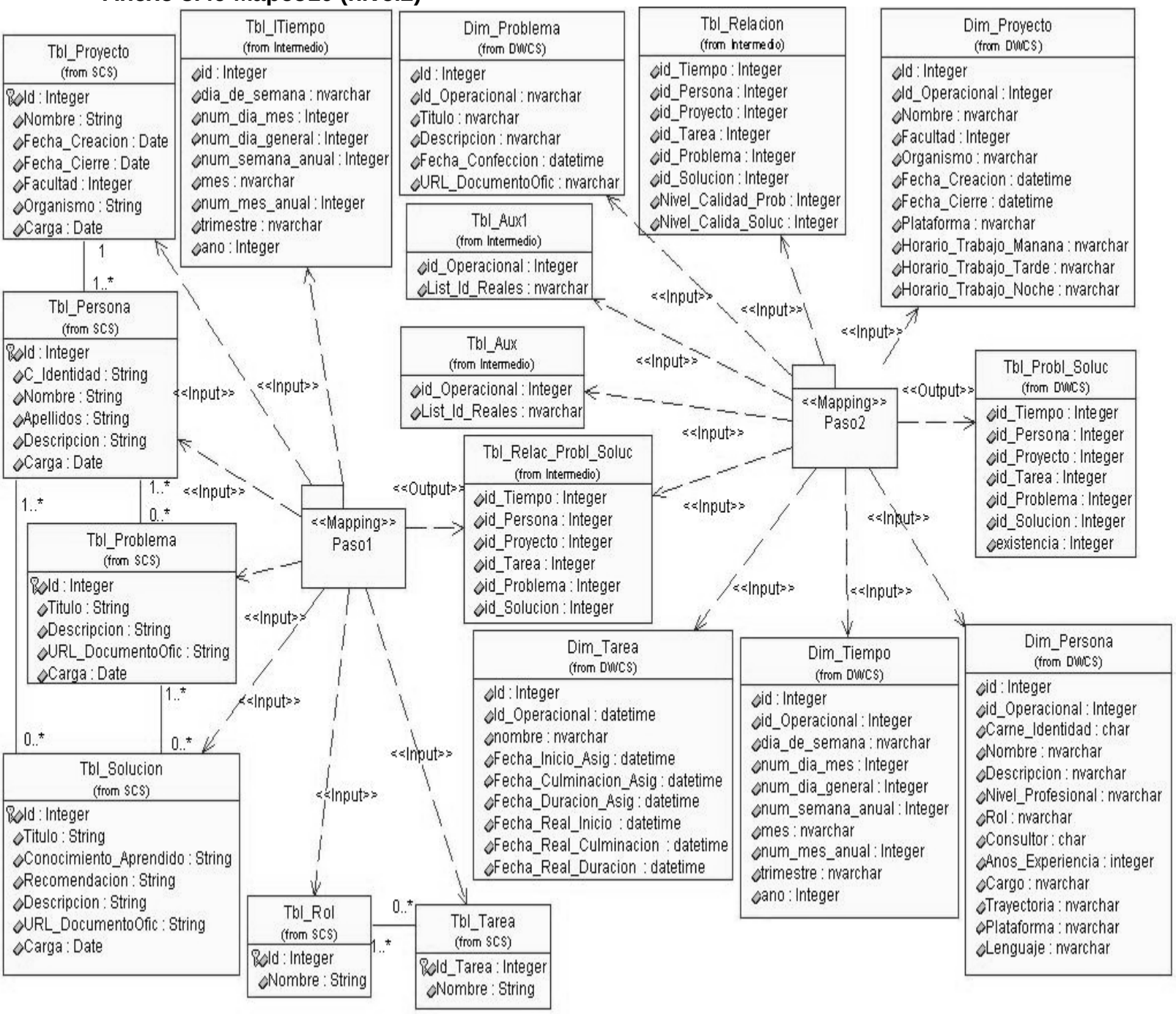
Anexo 3.47 Mapeo9 (nivel2)



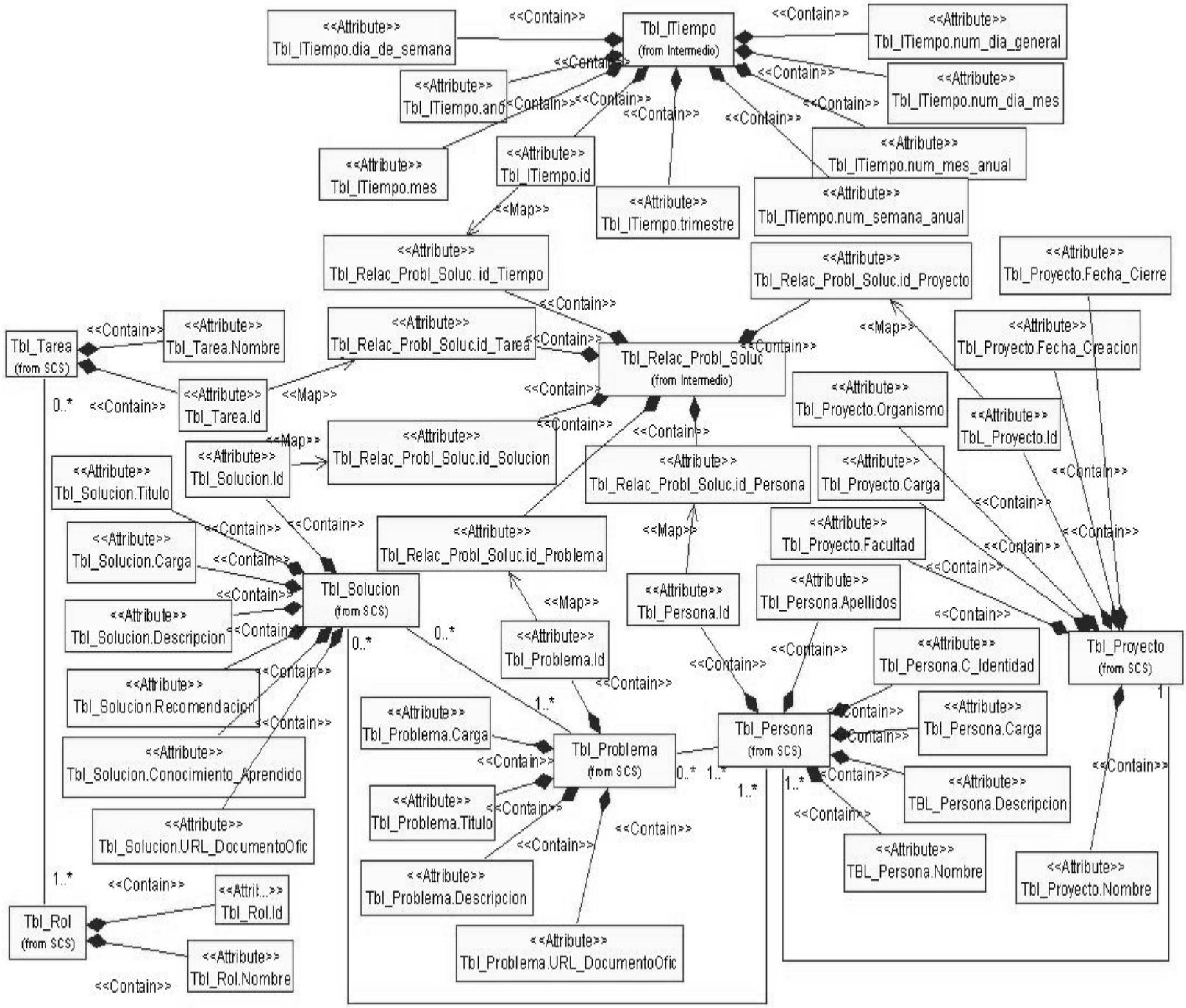
Anexo 3.48 Mapeo9. Paso1 (nivel3)



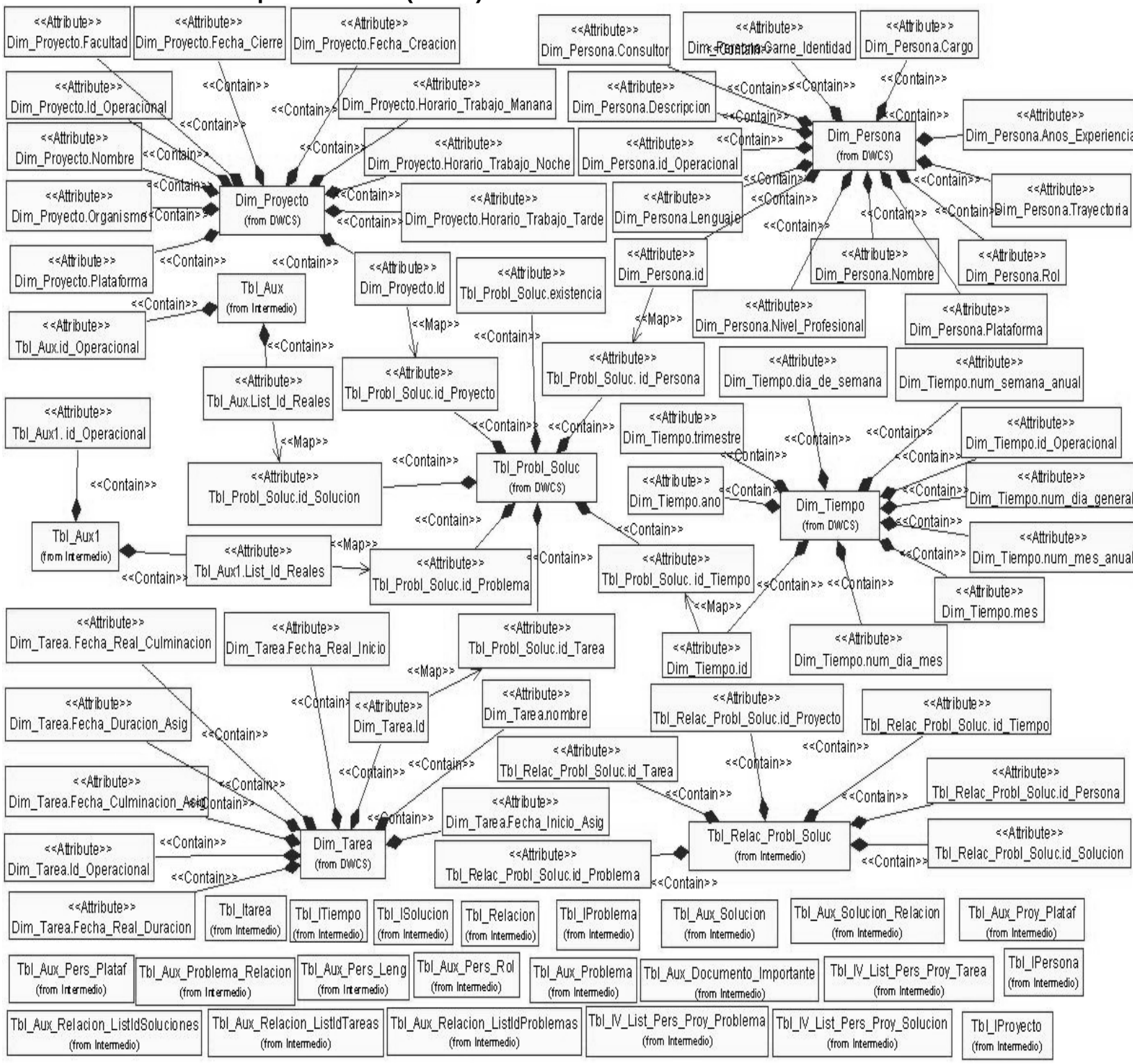
Anexo 3.49 Mapeo10 (nivel2)



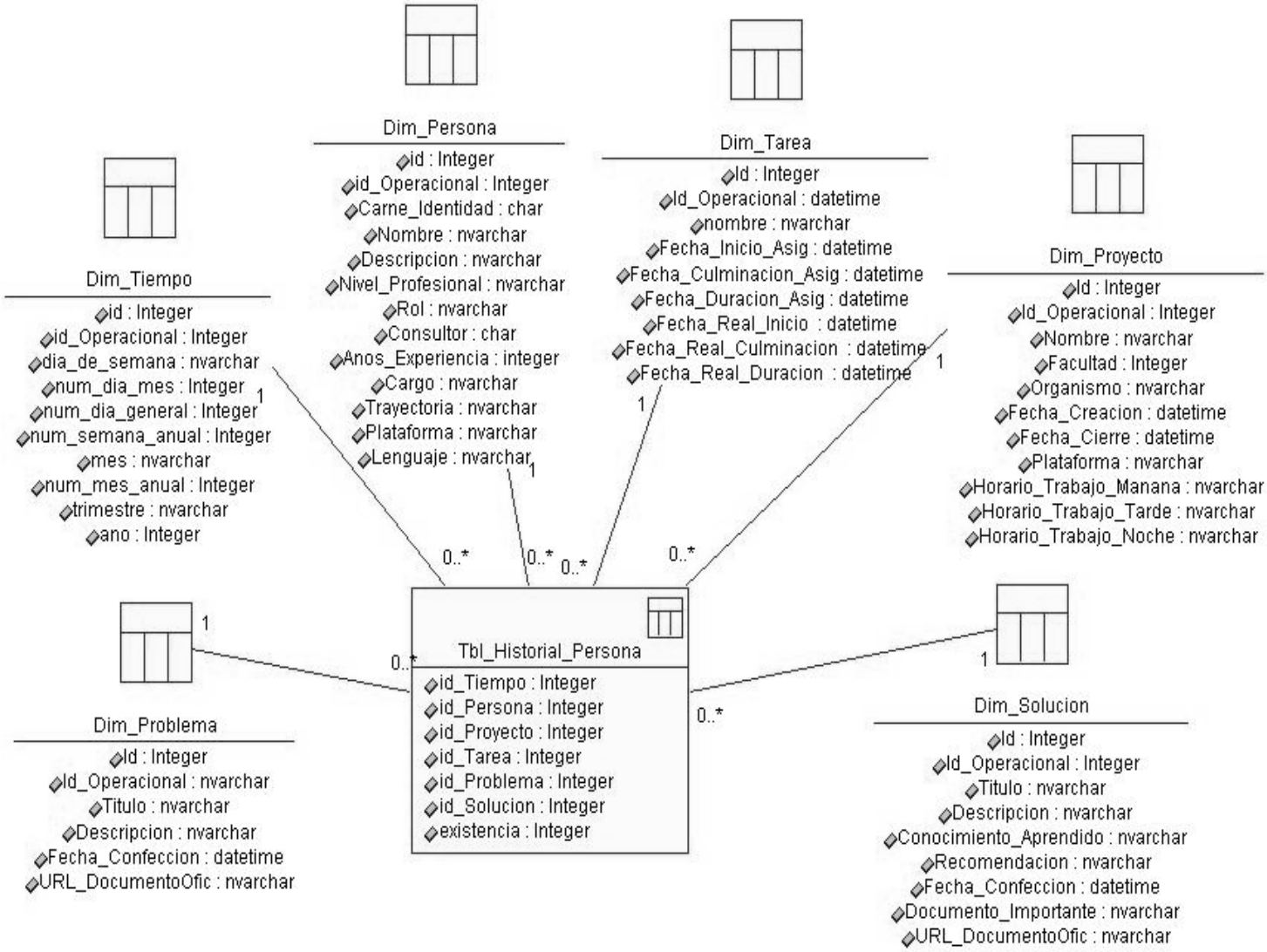
Anexo 3.50 Mapeo10. Paso1 (nivel3)



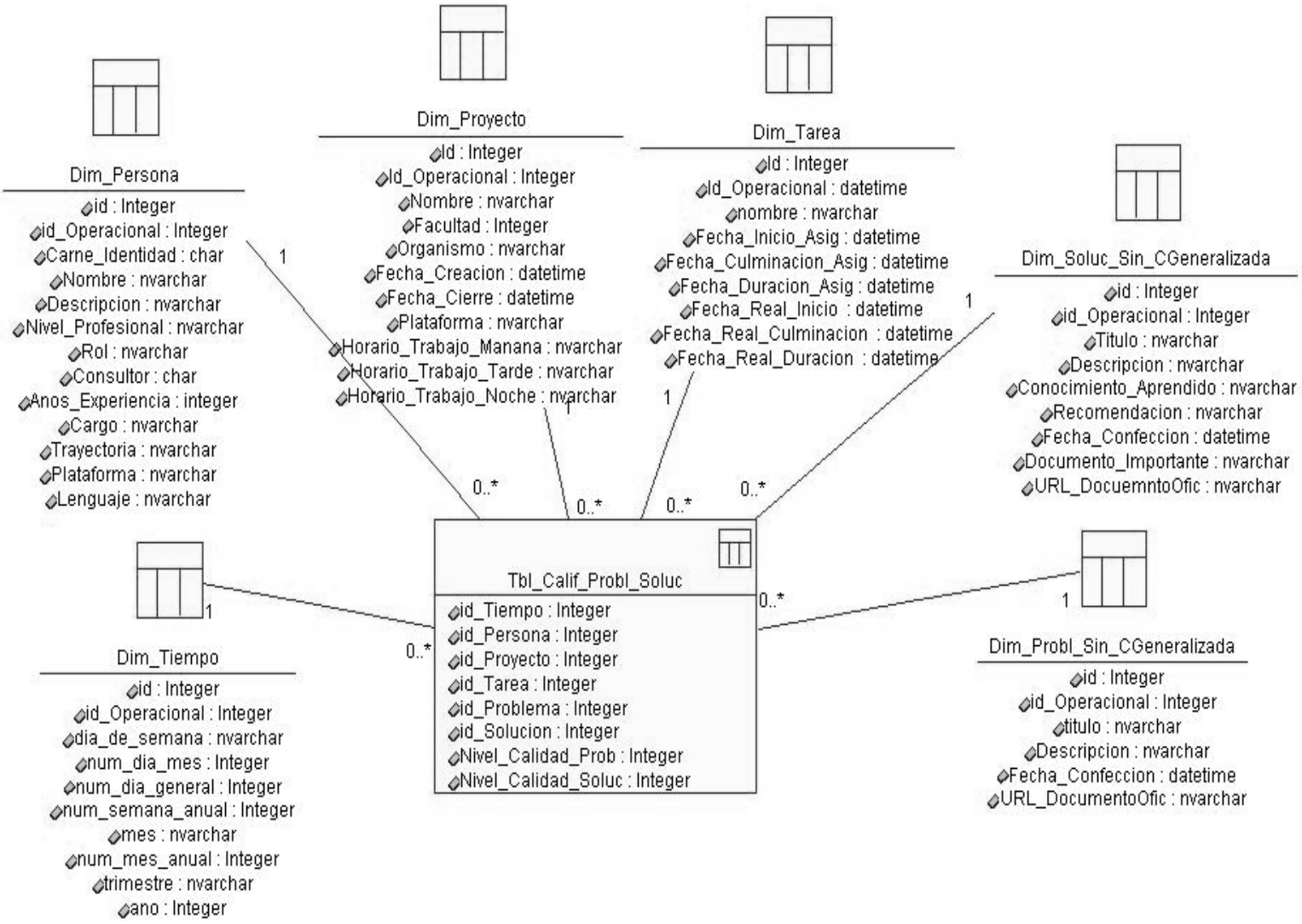
Anexo 3.51 Mapeo10. Paso2 (nivel3)



Anexo 3.52 Esquema lógico de DWH. Cubo de dato Historial de Persona.



Anexo 3.53 Esquema lógico de DWH. Cubo de dato Calificación de Problemas y Soluciones.



8 Glosario

A

Agregación: Tabla o estructura que contiene datos precalculados de un cubo.

B

Base de datos: Colección de datos -estructurada y organizada- para permitir el rápido acceso a la información de interés.

C

Conocimiento explícito: Conocimiento que puede ser estructurado, almacenado y distribuido.

Conocimiento tácito: Conocimiento que forma parte de las experiencias de aprendizaje personales de cada individuo y que, por tanto, resulta sumamente complicado, si no imposible, de estructurar, almacenar en repositorios y distribuir. Las tecnologías de la información y la comunicación sólo permitirán almacenar y distribuir conocimiento explícito.

H

Herramienta: Programa o aplicación usada principalmente para crear, manipular, modificar o analizar otros programas.

I

Inteligencia artificial: Programas diseñados para que su funcionamiento imite los procesos humanos de toma de decisiones y para que aprenda de los eventos pasados.

M

Mapeo: Proceso de convertir los datos que son transmitidos en un formato por el remitente, al formato de datos que puede ser aceptado por el receptor.

Minería de datos: Proceso de extracción de información y patrones de comportamiento que permanecen ocultos entre grandes cantidades de información.

P

Plataforma: Basamento, ya sea de hardware o software, sobre el cual un programa puede ejecutarse.

Protocolo: Descripción formal de formatos de mensaje y de reglas que dos ordenadores deben seguir para intercambiar dichos mensajes.

R

Repositorio: Sitio centralizado donde se almacena y mantiene información digital, habitualmente bases de datos o archivos informáticos.

S

Sistemas informacionales: Sistemas que permiten estudiar el comportamiento de la empresa, se utilizan para administrar y controlar la empresa

Sistemas operacionales: Sistemas que garantiza la automatización de los procesos y el flujo de la información a través de la organización, se utilizan en el funcionamiento de los negocio en tiempo real, basándose en datos actuales.

T

Tablespace: Unidad lógica de almacenamiento de datos representada físicamente por uno o más archivos de datos.

Toma de decisiones: Conjunto de actividades intelectuales o cibernéticas que utilizando una cierta información disponible dan como resultado una ciertas acciones, todo ello en un contexto real y concreto.

Transacciones: Suceso externo que involucra el traslado de algo de valor entre dos o más entidades.