Universidad de las Ciencias Informáticas

Facultad 2

Título: Diseño del Repositorio de Datos de la Sala Situacional del SIGEP desde el punto de vista de Clasificación y Atención Integral

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autor: Malena Garcia Izquierdo

Tutores: Ing. Gueorgui Obregón Obregón

Ing. René Martínez Bravet

Mayo 2011

La Habana

DECLARACIÓN DE AUTORÍA

Declaro que soy la única autora de este trabajo titulado: Diseño del repositorio de datos de la Sala Situacional del SIGEP desde el punto de vista de Clasificación y Atención Integral y autorizo a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmo la presente a los ____ días del mes de _____ de 2011.

Malena Garcia Izquierdo

_____

Firma del autor

Ing. René Martínez Bravet

_____

Firma del tutor

Ing. Gueorgui Obregón Obregón

_____

Firma del tutor

AGRADECIMIENTOS

A mi mamá por estar siempre a mi lado en cada paso que he dado, certero o no. Por todo el cariño que me brinda, por ser ejemplo de esfuerzo y sacrificio. Porque gracias a ella me he formado como mujer.

A mi papá y Adelaida por su apoyo y comprensión.

A mi abuela y mis tías por nunca dejarme sola y siempre estar al pendiente de mí.

A Iliana y Juan que tanto me han apoyado y ayudado, así como a todos los familiares y vecinos que han estado al tanto de mis estudios y mi realización como profesional.

A mi prima Anniat por ser, aunque lejos, la hermana que nunca tuve. Por hacer lo posible y lo imposible por tratar siempre de ayudarme. Gracias por estar ahí para mí.

A mis amigos, viejos y nuevos: Diane, Maipu, Danae, Cary, Puchito, Wilfre, Daire, Yadira. Gracias por brindarme su amistad desinteresada y compartir conmigo tantas cosas buenas que jamás se borrarán de mi memoria pero sobre todo, gracias por tenerme tanta paciencia.

A mi tutor René por ayudarme tanto. Gracias por hacerme reír.

A Isora porque a pesar de no conocerme me ayudó muchísimo con la corrección de este trabajo.

A la profe Anabel por brindarme su tiempo.

A todos los que de una manera u otra me brindaron su ayuda, tiempo y conocimiento.

Malena

DEDICATORIA

A mi mamá y a mi hermano: Quisiera ser siempre el orgullo de sus vidas.

Malena

RESUMEN

La Universidad de las Ciencias Informáticas (UCI) en el 2006 tuvo la tarea de desarrollar el Sistema de Gestión Penitenciaria (SIGEP). La implantación del sistema y el correspondiente crecimiento de los datos generados trajeron como consecuencia que se viera afectado el rendimiento de la aplicación ya que no cuenta con una solución relacionada con la inteligencia del negocio.

El propósito fundamental de este trabajo es el diseño de un Data Mart sobre la sala situacional del SIGEP desde el punto de vista de Clasificación y Atención Integral para así dar soporte al proceso de toma de decisiones por parte del sistema penitenciario venezolano.

El Data Mart está basado en la metodología DWEP, utilizando el enfoque propuesto por Ralph Kimball así como la arquitectura de dos capas. Como gestor de base de datos se hace uso de Oracle 10g, proponiendo además, la Suite de Inteligencia de Negocio Pentaho para el proceso de extracción, transformación y carga de los datos.

El equipo de desarrollo tiene entre sus metas la creación de un Data Mart para cada módulo del sistema, de manera que se puedan integrar finalmente para la construcción de un Data Warehouse central que ayude al proceso de gestión de información y una acertada toma de decisiones.

ÍNDICE

INDICE DE FIGURAS

INTRODUCCIÓN

El concepto de BI, Inteligencia del Negocio, se remonta a octubre de 1958, cuando Hans Peter Luhn de IBM, escribió un artículo titulado *"A Business Intelligence System"* en el cual describía las características que debía tener un sistema de este tipo. "El objetivo básico de la Business Intelligence es apoyar de forma sostenible y continuada a las organizaciones para mejorar su competitividad, facilitando la información necesaria para la toma de decisiones". (1)

El progreso irreversible del mundo y las nuevas tecnologías de la informática y las comunicaciones han acentuado el papel que juegan actualmente la información y el conocimiento en el desarrollo de toda empresa, más pronunciadamente con la aplicación de la inteligencia del negocio en la Informática.

La información generada diariamente crece de manera vertiginosa y su correcta gestión es fundamental para la toma de decisiones, la entrada a nuevos mercados, análisis de las necesidades de los clientes, determinación de la rentabilidad de un producto, formular estrategias del negocio, entre otros factores imprescindibles para toda empresa.

La eficiente gestión de la información de manera que ayude al proceso de toma de decisiones, es clave para la supervivencia en un mercado dinámico y competitivo como el actual. Aprender a competir utilizandoesta información es primordial para el desarrollo de las empresas.

Si se parte de los sistemas de origen de una organización (bases de datos, ERPs, ficheros de texto…) el análisis de los datos se vería como algo sencillo, pero estos sistemas suelen presentar la información de manera estática a través de una serie de informes ya predefinidos, o sea, no permiten explotar los datos, navegar entre ellos o manipularlos desde otras perspectivas, etc. Aunque estos sistemas están concebidos para controlar los datos generados en la empresa y ejecutar las decisiones operativas que conducen las actividades básicas, no permiten obtener toda la información deseada.

Para que la información se pueda gestionar de manera eficiente, debe utilizarse un almacén de datos, (DW, del inglés Data Warehouse). Los DW proporcionan una herramienta para la toma de decisiones en cualquier área del negocio, brindando una visión global de los datos de una organización.

De igual manera podemos entender un Data Mart como un subconjunto de los datos de un DW que tiene como objetivo responder a un análisis determinado y con una población de usuarios específica.

Los DW son actualmente un punto de atención de las organizaciones y proveen un ambiente para que las empresas hagan un mejor uso de la información que es administrada por diversas aplicaciones operacionales.

Con la llegada de Hugo Rafael Chávez Frías a la presidencia de Venezuela comienza la atención a los privados de libertad de ese país. En el año 2004 el mandatario decreta la fase de Emergencia Penitenciaria, implementándose, en el año 2005, el diagnóstico penitenciario, lo cual permitió determinar cuántos, quiénes y en qué condiciones vivían los reclusos venezolanos.

La Universidad de las Ciencias Informáticas (UCI) en el 2006 tuvo la tarea de desarrollar el Sistema de Gestión Penitenciaria (SIGEP) como parte del proyecto Humanización Penitenciaria con el objetivo de contribuir al control operativo, administrativo y estratégico del sistema penitenciario venezolano y de esta manera garantizar el respeto a los derechos de los internos, su actividad de rehabilitación y reinserción en la sociedad.

“El SIGEP, en su concepción inicial, establecida en el “*Proyecto Técnico de Asesoría Especializada, Colaboración Médica Odontológica, Comunicación Institucional y Solución Tecnológica para apoyar la modernización del Sistema Penitenciario de la República Bolivariana de Venezuela*” de agosto de 2006, concibe una solución de software cuyo objetivo general es: desarrollar e implantar un sistema informático que soporte las decisiones estratégicas del Ministerio del Interior y Justicia y de la Dirección General de Custodia y Rehabilitación del

Recluso, que contribuya a garantizar el respeto a los derechos de los internos, su actividad de rehabilitación y reinserción en la sociedad.” (2)

Por su parte “Sala situacional es un subsistema que solo contaba con una definición general de su objetivo y no se encontraba estructurado por módulos dado que no se había determinado en el estudio preliminar sus funciones específicas. Durante la modelación del negocio se identificaron cuatro tipos de salidas de información para el apoyo a la gestión de la sala situacional que apuntaban a soportar la toma de decisiones operativas, tácticas y estratégicas. Estas salidas se agrupan en: información operativa, avisos, alertas e información histórica; sin embargo, no todas estas funciones son propias de la sala situacional ya que el análisis históricos de datos y el monitoreo de indicadores de riesgos (avisos y alertas) son propios del departamento de Gestión de Riesgos. La Sala Situacional fundamentalmente se dedica al monitoreo de la situación operativa y responder a situaciones de emergencia. El departamento de Gestión de Riesgos por otro lado se dedica al análisis de la información.” (2)

La implantación del sistema y el correspondiente crecimiento de los datos generados trajeron como consecuencia que se viera afectado el rendimiento de la aplicación, a mayor escala al incluir reportes que precisan la búsqueda en varias tablas o varios módulos al mismo tiempo. Con el objetivo de darle una solución a este problema se crearon vistas materializadas[1] para tratar de hacer más eficiente el rendimiento de la aplicación de manera que las consultas a la base de datos se realizaran de forma más rápida. Esta variante no fue fructífera ya que las vistas se actualizaban sólo una vez al día y al sistema penitenciario le interesaba que la información fuera operativa, o sea, que a la hora de ser consultada se tuvieran los datos en tiempo real, y con las vistas no se lograba este objetivo al estar éstas en muchas ocasiones desfasadas con respecto a los datos reales. Por otra parte en el momento de actualizar las vistas disminuía el rendimiento de la aplicación al demorarse la actualización de los datos.

[1] En un sistema de gestión de base de datos que siga el modelo relacional, una vista es una tabla virtual, que representa el resultado de una consulta. Siempre que se consulta o se actualiza una vista normal, el SGBD convierte estas operaciones en consultas o actualizaciones de las tablas usadas para definir la vista.

De igual manera al establecer los requisitos funcionales se pensaron una cantidad de reportes predefinidos lo que trajo consigo una traba para la correcta toma de decisiones por parte de la administración ya que siempre que surja la necesidad de hacer un nuevo reporte los desarrolladores son los únicos capaces de darle solución a este problema.

El equipo de trabajo del proyecto tiene entre sus metas la creación de un Data Mart para cada módulo del sistema, de manera que se puedan integrar finalmente para la construcción de un Data Warehouse central que ayude al proceso de gestión de información y una acertada toma de decisiones.

Por lo anteriormente expuesto se plantea como **problema a resolver**: ¿Cómo facilitar la toma de decisiones de la administración teniendo en cuenta la clasificación y la atención integral a la población penal?

Como **objeto de estudio** proceso de diseño de un Data Warehouse y como **objetivo general** de la investigación: Diseñar un Data Mart orientado a la información referente a la población penal clasificada y atendida para la toma de decisiones en la Sala Situacional del SIGEP.

Por lo que se especifica el siguiente **campo de acción**: Proceso de diseño de un Data Mart para la población penal clasificada y atendida de la Sala Situacional del SIGEP.

Determinado el objetivo general lo podemos dividir en los siguientes **objetivos específicos**:

1. Elaborar el marco teórico de la investigación
2. Determinar los requisitos del sistema.
3. Definir la arquitectura del sistema.
4. Diseñar la Base de Datos del Data Mart.
5. Diseñar e implementar los procesos de Extracción, Transformación y Carga (ETL).

Como **posibles resultados** de la investigación se encuentran el diseño de los cubos de datos asociados al Data Mart y la propuesta de los procesos ETL necesarios para poblar los cubos.

El trabajo está estructurado de la siguiente manera: resumen, dos capítulos de contenido, conclusiones, recomendaciones, referencias bibliográficas y anexos.

En el capítulo 1 (Fundamentación Teórica) se recogen conceptos, objetivos y características de los Data Warehouse, las tecnologías y software empleados en su diseño, así como un análisis de las diferencias y funcionalidades de los principales gestores de bases de datos que soportan Data Warehouse.

En el capítulo 2 (Diseño del Data Mart) contiene el diseño del Data Mart e implementación de los procesos de Extracción, Transformación y Carga, presentando además los fundamentos de la metodología a utilizar.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

1. Introducción

En el presente capítulo se realiza un análisis de conceptos, objetivos y características de los Data Warehouse, las tecnologías y software empleado en su diseño. Se evidencian también las diferencias y funcionalidades de los principales gestores de bases de datos, para la exploración de datos y la extracción de conocimiento.

1.1. Sistema de Gestión Penitenciaria (SIGEP)

El SIGEP desarrollado en la UCI y desplegado en Venezuela tiene como objetivo “desarrollar e implantar un sistema informático que soporte las decisiones estratégicas del Ministerio del Interior y Justicia y de la Dirección General de Custodia y Rehabilitación del Recluso, que contribuya a garantizar el respeto a los derechos de los internos, su actividad de rehabilitación y reinserción en la sociedad.” (2)

El subsistema *Observación, Clasificación y Tratamiento* permite controlar las actividades de reinserción de los privados de libertad. Dentro de este subsistema se encuentran definidos los módulos que atañen al diseño del Data Mart en cuestión.

Clasificación: Finalizado el período de observación, el individuo es diagnosticado y un equipo multidisciplinario lo clasifica en alguno de los regímenes de seguridad penal (máxima, media o mínima), para ofrecerle un tratamiento individualizado. El módulo permitirá registrar la clasificación dada a cada individuo, la fecha en que se le otorgó así como mantener un histórico de las clasificaciones que ha tenido y los motivos por los cuales ha cambiado de una a otra.

Educación: El módulo permite registrar los distintos tipos de actividades educativas (formales y no formales) que se desarrollan en los Establecimientos Penitenciarios. Las actividades formales se refieren a las que elevan el nivel de escolaridad de los privados de libertad (primaria, secundaria, bachillerato, universidad). Las no formales se dedican a la formación de oficios, manualidades, etc. Además el módulo gestiona la matrícula de las diferentes actividades

educativas, permite mantener un registro actualizado de la asistencia de los internos a las actividades y registrar las evaluaciones finales de cada curso.

En el expediente penitenciario de cada privado de libertad queda reflejado un historial de cursos matriculados, la evaluación final obtenido y el total de horas dedicadas al estudio.

Trabajo: Este módulo permite controlar las Unidades Productivas que se conforman en los Establecimientos Penitenciarios, la incorporación de los internos a ellas, así como el registro de la asistencia diaria a las actividades laborales. En el expediente penitenciario de cada privado de libertad aparece un resumen del total de horas laboradas, así como las unidades productivas en las que ha estado incorporado.

Deporte y cultura: Con este módulo se controla la incorporación y participación de los privados de libertad en las manifestaciones culturales o disciplinas deportivas que se practican en el penal, bajo la orientación y supervisión de las coordinaciones de deporte y cultura.

El SIGEP no incluía, entre sus versiones desplegadas o en desarrollo, una funcionalidad relacionada con la inteligencia del negocio por lo que se decidió poner en práctica esta solución en el proyecto empezando por el subsistema Control Penal de la Sala Situacional del SIGEP. Este utiliza como herramienta de inteligencia del negocio a la Suite de Business Intelligence Pentaho, sobre un almacén de datos dimensional para la generación de reportes de manera eficiente.

Esta solución es la estudiada y analizada en la presente investigación y la que se tomará como punto de partida para el diseño del Data Mart.

1.2. Definición de Data Warehouse

Actualmente las empresas necesitan depositar toda su confianza en la toma de decisiones, por lo que se requiere de información consistente, integrada, histórica y preparada para ser analizada. La aparición de los Data Warehouse o Almacenes de Datos es la respuesta a estas necesidades. Para poder entender a profundidad el funcionamiento de un Data Warehouse se

hace imprescindible citar a dos destacados pioneros en este tema: William Harvey Inmon o Bill Inmon, reconocido por muchos como el padre de los Data Warehouse y Ralph Kimball, considerado como el promotor del modelado dimensional para el diseño de almacenes de datos.

De manera general podemos entender un Data Warehouse como una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta.

Los Data Mart son un subconjunto de los datos de un Data Warehouse y están dirigidos a un área específica de la empresa como pueden ser Finanzas y Recursos. Ralph Kimball lo define como "una restricción de un Data Warehouse para un determinado proceso de la empresa y que debe seguir los requisitos generales de implementación que posee un Data Warehouse." (3)

1.3. Características y objetivos de un Data Warehouse

A decir de Inmon un Data Warehouse "…es una colección de datos orientada a temas, integrada, no volátil e indexado en el tiempo para dar soporte a la toma de decisiones." (4)

De esta manera Inmon establece las características de los Data Warehouse, explicadas a continuación:

Orientada a temas: Cada parte del Data Warehouse debe estar orientada a resolver un problema de la empresa. Los sistemas orientados a las aplicaciones contienen datos para satisfacer requerimientos funcionales o de procesamiento los cuales pueden ser usados o no a la hora de la toma de decisiones, mientras que en los almacenes de datos se obvia toda la información que no se necesitará para la misma.

A diferencia de los ambientes operacionales donde las estructuras de datos se diseñan en función de las aplicaciones, en los almacenes de datos se diseña sobre las principales áreas de la empresa, tales como, ventas, clientes, productos, etc. Esto facilita su acceso y respuesta a las

peticiones ya que toda la información referente a un área determinada del negocio residirá en el mismo lugar.

Figura 1: Ejemplo de organización de los datos orientados a temas

Integrada: De todos los aspectos de un Data Warehouse, la integración de los datos es lo más importante. La información debe ser transformada en medidas, códigos y formatos comunes para que pueda ser útil. La integración permite a la empresa implementar la estandarización de sus definiciones.

En la Figura 2 se evidencia cómo se lleva a cabo el proceso de integración en un almacén de datos, llevando las distintas codificaciones existentes en el ambiente operacional a una sola en el Data Warehouse.

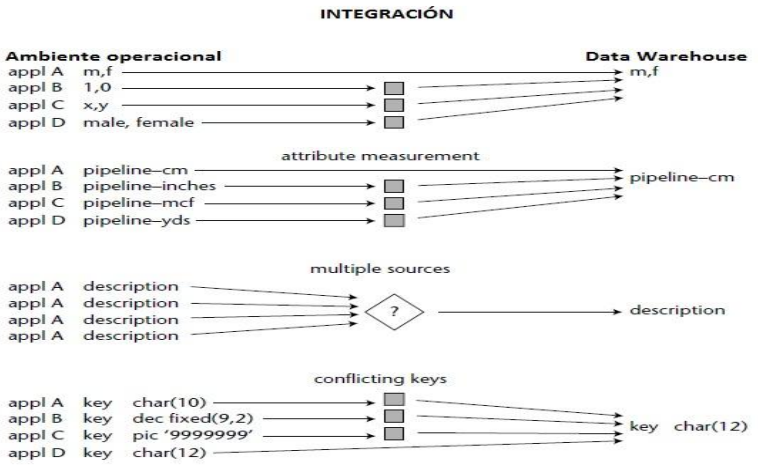


Figura 2: Ejemplo de integración

No volátil: La tercera característica importante de un Data Warehouse es que no sea volátil. En los ambientes operacionales la información puede ser manipulada, ya sea, insertar, modificar, eliminar. Mientras que la información contenida en un Data Warehouse puede ser leída pero no modificada. De esta manera la información es permanente. La actualización del Data Warehouse significa la incorporación de los valores que tomaron las variables contenidas en él sin realizar ninguna acción sobre lo que ya existía.

En la figura 3 se muestra cómo los datos son accedidos y modificados en el ambiente operacional a diferencia de los depósitos de datos donde es los datos son cargados y accedidos pero no modificados.

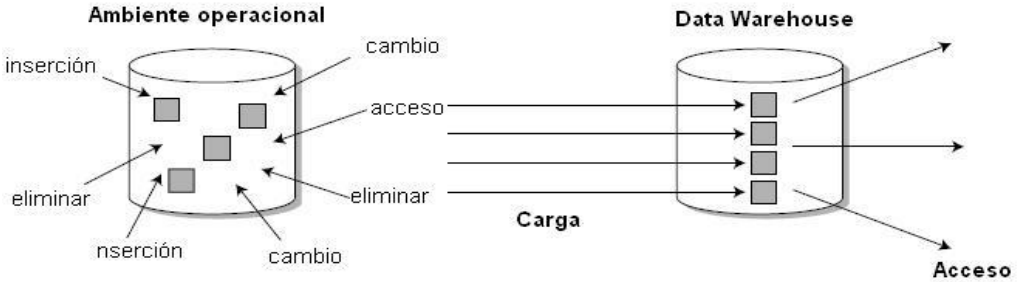


Figura 3: Ejemplo de no volatilidad

Indexado en el tiempo: En las bases de datos operacionales los datos reflejan el estado de la actividad de la empresa en el momento presente y son actualizados según las necesidades que surjan. En los Data Warehouse la información se mantiene almacenada históricamente los que permite realizar comparaciones, análisis de tendencias y previsiones.

La figura 4 muestra el contraste de esta característica entre el ambiente operacional y un Data Warehouse.

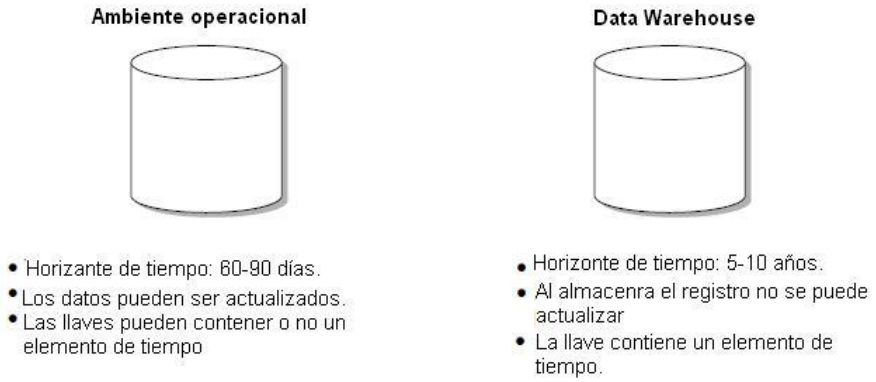


Figura 4: Ejemplo de característica de variación de tiempo

En un Data Warehouse se pueden establecer como objetivos más importantes los siguientes:

- Brindar una visión única y global de los clientes de la empresa.
- Aumentar la productividad.
- Mejorar la capacidad de respuesta a los problemas.
- Monitorear el comportamiento de los clientes.
- Predecir la compra de determinados productos.
- Disminuir el tiempo de espera a la hora de emitir un informe.
- Brindar la mayor cantidad de información posible a la mayor cantidad de usuarios posibles.

1.4. Modelos más utilizados en la construcción de Data Warehouse

Entre los métodos más utilizados en la actualidad para el diseño de un Data Warehouse se encuentran los propuestos por Inmon y Kimball. Ambos centran sus propuestas en los datos en sí mismos, estando de acuerdo en que un Data Mart o un Data Warehouse independiente no satisfacen las necesidades que tienen las empresas de acceder a la información de manera rápida y precisa y centrándose siempre en la construcción de una arquitectura robusta que se adapte a los cambios de las necesidades del negocio. Sus propuestas difieren en cuanto al modelo de datos y a la arquitectura propuesta.

Para resolver el problema de que en ocasiones los usuarios no necesitan tener acceso a toda la información de la empresa sino a una parte de esta se crean los Data Mart.

1.4.1. Modelo Dimensional

El Modelo Dimensional, propuesto por Kimball, es uno de los más utilizados en la construcción de un Data Warehouse. “Este está constituido por modelos de tablas y relaciones con el fin de mejorar el proceso de toma de decisiones de la empresa. Cada modelo dimensional está conformado por una única tabla en el esquema, llamada tabla de hechos, con múltiples enlaces conectándola a otras tablas llamadas dimensiones.” (5)

Podemos definir los elementos de estas tablas como:

Hechos: Colección de piezas de datos o datos de contexto. Cada hecho va a representar una parte del negocio, una transacción o un evento.

Dimensiones: Colección de miembros, unidades o individuos del mismo tipo.

Medidas: Son atributos numéricos de un hecho que representan el comportamiento del negocio relativo a una dimensión.

Se caracteriza por ser sencillo de crear, estable en presencia de cambios, además de mostrarse muy intuitivo y comprensible.

Kimball resalta que “los Data Marts están basados en los datos de la fuente y no en la visión departamental.” (6) 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. Con la utilización de esta estrategia se debe tener claro el plan de acción, es decir, qué áreas cubriremos y la integración de los distintos modelos.

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 y que cubran las necesidades de la organización, para luego de tenerlos desarrollados y funcionales se comience con la construcción del Data Warehouse basado en los datos que estos contienen.

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.

También William H. Inmon reconoce al Modelo Dimensional como bueno para el desarrollo de los Data Marts por las ventajas brindadas, pero propone la construcción del Data Warehouse basado en el Modelo Entidad Relación.

1.4.2. Modelo relacional

La idea defendida por Inmon sugiere la utilización de un modelo relacional, proponiendo que el modelo entidad relación es mucho más flexible que el dimensional. (4)

En cuanto a la arquitectura la idea defendida por Inmon plantea que “la construcción del Data Warehouse no debe ser sustituida por la implementación de varios Data Marts.” (4)

O sea, propone definir un Data Warehouse corporativo y a partir de este construir los modelos de análisis para los distintos niveles de la empresa. Con esta metodología los Data Marts se crearán después de haber terminado el Data Warehouse completo de la organización. Al ver de Inmon la sustitución de una Data Warehouse por Data Marts trae desventajas ya que los mismos estás diseñados para un área específica de la empresa, lo que puede llevar a diferencias entre las estructuras de datos contenidos en ellos, que al integrarlos al Data Warehouse algunos dejarán de ser flexibles y reusables.

Inmon resalta que en el proceso de construcción, el Data Warehouse se forma de los sistemas operacionales existentes, creándose áreas de diferentes temas. Cuando exista una cierta cantidad de éstas, 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.

Se decide utilizar el enfoque de Raplh Kimball para el diseño del Data Mart, pues es menos costoso, más funcional, se pueden priorizar las áreas críticas; su estructura de datos ofrece mayor facilidad al usuario para la exploración y búsqueda de información. Además en el proyecto ya existe implementado un Data Mart para el Control Penal por lo tanto se seguirán haciendo Data Mart a cada módulo del proyecto y luego se unirán en un almacén central. Igualmente se utilizará el modelo dimensional como modelo a seguir para diseñar el almacén de datos puesto que propone una forma sencilla de representar los datos y mejora el tiempo de consulta en la base de datos.

1.5. Arquitectura conceptual de los datos

“La estructura que reúne a todos los componentes de un Data Warehouse es conocida como arquitectura. La arquitectura incluye todo lo necesario para preparar los datos y almacenarlos.

Está compuesta por reglas, procedimientos y funciones que permiten al Data Warehouse trabajar y cumplir con los requerimientos de la empresa. Define normas, medidas, diseño general y técnicas de apoyo.” (7)

Usualmente suelen representarse varias capas en un Data Warehouse, a través de las cuales circulan los datos, nombrándose de acuerdo al número de capas que abarcan, de modo que los datos de una capa se obtienen a partir de los datos de la capa previa.

En la arquitectura de datos de **Una Sola Capa,** la información se guarda sólo una vez en el Data Warehouse, de manera que se almacenan solamente los datos de tiempo real, sobre los cuales actúan sistemas informacionales y operacionales.Esto puede traer contradicciones ya que ambos sistemas actúan sobre el mismo conjunto de datos y tal vez en el momento en que sean necesarios no estén disponibles para los fines operacionales porque pueden estar siendo consultados y mientras esto sucede no es posible realizar actualizaciones. Ver Figura 5.

Figura 5: Arquitectura de datos de una sola capa

En la arquitectura de datos de **Dos Capas,** se perfecciona la arquitectura de una sola capa, 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, estos 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. Ver Figura 6.

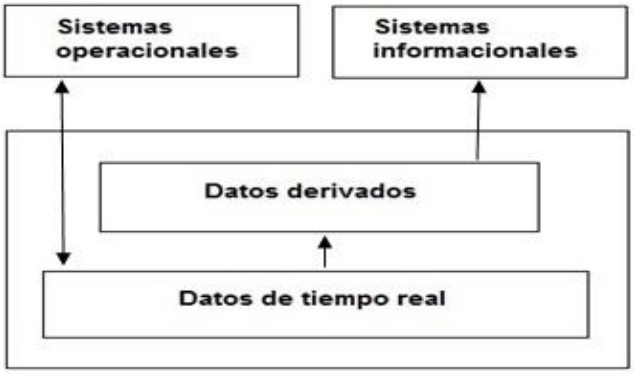


Figura 6: Arquitectura de datos de 2 capas

La arquitectura de datos de **Tres Capas** requiere una capa intermedia para llevar a cabo la transformación de los datos de tiempo real a datos derivados y así solucionar los problemas de inconsistencia, realizando el procesamiento de los distintos conjuntos de datos de tiempo real. A esta nueva capa se le conoce como capa de datos reconciliados. Figura 7.

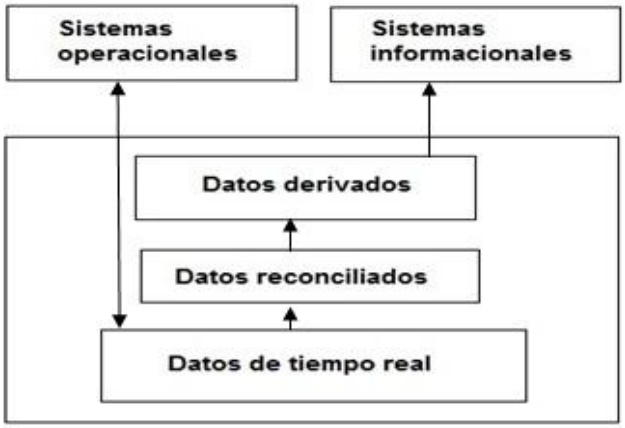


Figura 7: Arquitectura de datos de 3 capas

De acuerdo a las necesidades y

características de la empresa se podrá seleccionar cualquiera de las arquitecturas de datos anteriores.

1.6. Arquitectura conceptual de un Data Warehouse.

Para comprender cómo se relacionan todos sus componentes de un Data Warehouse es necesario contar con un modelo de Arquitectura Data Warehouse. Considerando su estructura en un marco de ocho niveles:

Nivel operacional: Se definen las diversas bases de datos operacionales y fuentes externas de donde serán extraídos.

Nivel de acceso a la información: Este es el nivel del que el usuario final se hace cargo directamente. Incluye las herramientas de consultas, análisis, generadores de informes y herramientas de data mining, (en español Minería de datos) que es la básicamente se encarga de extraer la información no trivial que está implícita, teniendo como finalidad la manipulación, análisis y presentación de los datos de acuerdo a los requerimientos de los usuarios. Su finalidad es servir de soporte a las decisiones gerenciales.

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. Con este nivel se logra que el usuario, sin tener en cuenta la herramienta de acceso a la información o la ubicación, sea capaz de acceder a todos los datos de la empresa necesarios

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. 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.

Los **metadatos** son básicamente datos acerca de los datos contenidos en el Data Warehouse. Así, uno de los problemas con que pueden encontrarse los usuarios es saber lo que hay en él y cómo puede acceder a lo que quieren.

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.

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, además, de controlar varios procesos con el propósito de conservar actualizado el Data Warehouse.

El Nivel de mensaje de la aplicación: Tiene que ver con la 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.

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 puedeverse el Data Warehouse simplemente como una vista lógica, pues puede no involucrar almacenamiento de datos.

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.

1.7. Organización física

ROLAP (Relational On Line Analytic Processing): Se implementa sobre tecnología relacional, dispone de algunas facilidades para mejorar el rendimiento.Aunque el almacén de datos se organiza como una base de datos multidimensional es soportado por un SGBD Relacional, las dimensiones y tablas de hechos serán traducidas a tablas.

MOLAP (Multidimentional On Line Analytic Processing): Almacena físicamente los datos en estructuras multidimensionales de manera que la representación externa y la interna coinciden.La tecnología está optimizada para consultas y análisis.No es adecuada para muchas dimensiones ni muchos registros.

HOLAP (Hybrid On Line Analytic Processing): Constituye un sistema híbrido entre MOLAP y ROLAP pues combina estas dos implementaciones para almacenar algunos datos en un motor relacional y otros en una base de datos multidimensional.

1.8. On-Line Transactional Processing (OLTP) y On-Line Analytical Processing (OLAP)

"Los **sistemas OLTP** son bases de datos orientadas al procesamiento de transacciones. Una transacción genera un proceso atómico (que debe ser validado con un *commit*, o invalidado con un *rollback*), y que puede involucrar operaciones de inserción, modificación y borrado de datos. El proceso transaccional es típico de las bases de datos operacionales." (8)

- El acceso a los datos está optimizado para tareas frecuentes de lectura y escritura. (Por ejemplo, la enorme cantidad de transacciones que tienen que soportar las BD de bancos o hipermercados diariamente).
- Los datos se estructuran según el nivel aplicación (programa de gestión a medida, ERP o CRM implantado, sistema de información departamental).

- Los formatos de los datos no son necesariamente uniformes en los diferentes departamentos (es común la falta de compatibilidad y la existencia de islas de datos).
- El historial de datos suele limitarse a los datos actuales o recientes.

"Los **sistemas OLAP** son bases de datos orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejo etc. Este sistema es típico de los Data Marts." (8)

- El acceso a los datos suele ser de sólo lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o eliminaciones.
- Los datos se estructuran según las áreas de negocio, y los formatos de los datos están integrados de manera uniforme en toda la organización.
- El historial de datos es a largo plazo, normalmente de dos a cinco años.
- Las bases de datos OLAP se suelen alimentar de información procedente de los sistemas operacionales existentes, mediante un proceso de extracción, transformación y carga (ETL).

1.9. Modelado de datos.

Existen tres niveles de modelado de datos: conceptual, lógico y físico. Las diferencias entre los almacenes de datos con las bases de datos operacionales en cuanto al tipo de consultas y rendimiento esperado, hacen que lasestrategias de diseño y los modelos de datos utilizados para el almacén de datos sean diferentes.

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.

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 foráneas, sin tener en cuenta cómo ellos se implementarán físicamente en la base de datos. Es un puente entre el nivel conceptual y el físico.

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

1.10. Metodología a utilizar

La finalidad de una metodología es guiar todo el proceso de desarrollo de un software, y en nuestro caso, de un almacén de datos. Las principales metodologías que se tendrán en cuenta en esta investigación son Hefesto y DWEP.

La metodología **Hefesto** es independiente de las herramientas de implementación utilizadas, de las estructuras físicas que contengan el almacén de datos y del tipo de vida empleado para contener a la metodología. Los modelos utilizados, conceptuales y lógicos, son sencillos de crear y analizar. Su estructura se adapta fácilmente a los cambios del negocio ya que está basada en los requerimientos del usuario.

“Esta metodología comienza recolectando las necesidades de información de los usuarios y consiguiendo las preguntas claves para la empresa. De esta manera se obtienen indicadores y perspectivas de análisis que serán de gran utilidad para la construcción del modelo conceptual de datos del almacén de datos. Acto seguido se lleva a cabo el análisis de los OLTP para marcar las correspondencias con los datos fuentes y seleccionar los campos de estudio para cada perspectiva. Luego se pasa a la construcción del modelo lógico del almacén de datos y finalmente se definirán los procesos de extracción, transformación y carga así como la limpieza de los datos fuente.” (9)

Debido a una gran variedad de modelos utilizados en las fases de diseño de los almacenes de datos se desarrolló una metodología 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. (10)

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, y está basada en el Lenguaje Unificado de Modelado (UML por sus siglas en inglés) y el Proceso Unificado de Desarrollo de Software (RUP por sus siglas en Inglés) (11). El que esté basado en UML y RUP evita que los diseñadores aprendan una nueva notación o lenguaje para el diseño de almacenes de datos.” (10)

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
- Lógico
- Físico

Como el DWEP es una instanciación de RUP para el desarrollo de almacenes de datos, establece, al igual que esta, “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. Además adiciona dos nuevas actividades: Mantenimiento y Revisión Pos-desarrollo.” (10) En cada flujo de trabajo se utilizan diferentes diagramas UML, para modelar y documentar el proceso de desarrollo.

Para el diseño del Data Mart se desecha la opción de la utilización de la metodología Hefesto ya que su principal ventaja consiste en ir analizando con los clientes, en cada flujo de trabajo, el avance en el desarrollo del almacén de datos, para así ir agregando nuevos requisitos si es necesario, cosa que no será posible en este caso ya que el mismo es una propuesta que se le hará a la dirección del sistema penitenciario venezolano una vez implementado el data Warehouse completo. Se utilizará la metodología 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 entre los mismos 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 Visual Paradigm en su versión 3.4 para representar cada uno de los modelos de la metodología a utilizar.

Visual Paradigm para UML es una herramienta UML profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. El software de modelado UML ayuda a una rápida construcción de aplicaciones de calidad, mejores y a un menor costo. Permite dibujar todos los tipos de diagramas de clases, ingeniería inversa, generar código desde diagramas y generar documentación. Es multiplataforma, muy fácil de usar y con un ambiente gráfico agradable para el usuario, además permite modelar base de datos y transformación de diagramas de Entidad-Relación en tablas de base de datos, posibilitando la creación y diseño del modelo de datos.

1.11. Herramientas de creación de almacenes de datos

Para llevar a cabo la construcción de un almacén de datos es de vital importancia contar con una serie de herramientas que faciliten el trabajo. Cuando estas herramientas contienen una suite completa que permite desde la creación de las base de datos hasta la explotación de la misma para diferentes perfiles y objetos, son conocidas como herramientas de Business Intelligence.

El Data Warehouse siempre se va a implementar sobre un sistema gestor de base de datos (SGBD).

Un SGBD muy conocido en el mundo es **Oracle**, fabricado por Oracle Corporation. Se considera uno de los sistemas de bases de datos más completos. Es un sistema gestor de base de datos líder en la industria, utilizable para almacenar todo tipo de datos, incluyendo datos relacionales, documentos, multimedia, XML y datos de localización. Es un sistema robusto, tiene muchas características que garantizan la seguridad e integridad de los datos; que las transacciones se ejecuten de forma correcta, sin causar inconsistencias.

Ayuda a administrar y almacenar grandes volúmenes de datos. Es estable y escalable, fiable. 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.

Teniendo en cuenta sus características se utilizará Oracle 10g pues es capaz de manejar eficientemente todo el volumen de información del Data Mart y proporciona todo el soporte necesario para la construcción y mantenimiento del mismo, además de ser el gestor de base de datos utilizado en el proyecto donde en un futuro será implantado el Data Mart.

La principal desventaja de la utilización de Oracle como herramienta son los elevados precios de las licencias de software y del soporte técnico, lo que lo hace un gestor típico de sistemas informáticos de grandes compañías o instituciones gubernamentales.

Oracle provee una potente herramienta para el diseño de Data Warehouse llamada **Oracle Warehouse Builder (OWB)** la cual brinda una herramienta de diseño fácil de utilizar, de integración empresarial, que administra todo el ciclo de vida de los datos y metadatos para Oracle Database 10*g*.

Permite 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 almacenes de datos. Es una herramienta orientada no solamente a realizar el proceso de ETL, sino también la definición, administración y mantenimiento de un Data Warehouse.

Incorpora 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.

Permite crear el diseño lógico describiendo los cubos OLAP en dimensiones, jerarquías, medidas, medidas pre-calculadas y cualquier otro componente que se necesite. Con el uso del 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 o en ROLAP.

A pesar de que OWB es una gran herramienta potente para la creación de almacenes de datos, no será la utilizada ya que el cliente no pagó la licencia de la misma.

La plataforma Open Source **Pentaho Business Intelligence** cubre muy amplias necesidades de Análisis de los Datos y de los Informes empresariales, lo que reduce enormemente los costos de implantación del Data Warehouse, aspecto que constituye uno de los principales problemas a la hora de decidirse a crear un almacén de datos.

“Las soluciones de esta plataforma están escritas en Java y tienen un ambiente de implementación también basado en Java. Eso hace que Pentaho sea una solución muy flexible para cubrir una amplia gama de necesidades empresariales, tanto las típicas como las sofisticadas y especificas al negocio” (12). La Suite Pentaho se define a sí mismo como una plataforma de BI “orientada a la solución” y “centrada en procesos” que incluye todos los principales componentes requeridos para implementar soluciones basados en procesos.

Las soluciones que Pentaho pretende ofrecer se componen fundamentalmente de una infraestructura de herramientas de análisis e informes integrados con un motor de workflow (en español flujo de trabajo) de procesos de negocio. La plataforma será capaz de ejecutar las reglas de negocio necesarias, expresadas en forma de procesos y actividades y de presentar y entregar la información adecuada en el momento adecuado.

Pentaho está construido en torno al servidor de aplicaciones J2EE JBoss y Jboss Portal, antes de ser adquirida por Red Hat, habilitando que toda la información sea accesible mediante un navegador en la intranet de la empresa. El mismo presenta informes en los formatos habituales (html, xls, pdf...) mediante JfreeReport, u otras plataformas como BIRT o JasperReports. Para la generación de PDFs utilizan, como podría ser previsible, el conocidísimo Apache FOP.

Pentaho Analisis suministra a los usuarios un sistema avanzado de análisis de información. Con uso de las tablas dinámicas (pivottables, crosstabs), generadas por Mondrian y JPivot, el usuario puede navegar por los datos, ajustando la visión de los datos, los filtros de visualización, añadiendo o quitando los campos de agregación. Los datos pueden ser representados en una forma de SVG o Flash, los dashboards widgets, o también integrados con los sistemas de minería de datos y los portales web (portlets). Además, con el Microsoft Excel Analysis Services, se puede analizar los datos dinámicos en Microsoft Excel (usando la conexión a OLAP server Mondrian).

Los WebServices[2] (en español servicios web) son una característica fundamental de Pentaho. El mismo utiliza como motor de WebServices Apache Axis, quedando los servicios descritos en el lenguaje de definición de servicios web WSDL. Dos son los fundamentos del workflow de procesos de negocio: el motor de workflow Enhydra Shark y el estándar WPDL, auspiciado por la WorkFlow Management Coalition (WFMC), organismo que declara tener más de 300 empresas asociadas, incluyendo a las conocidas IBM, Oracle, BEA, Adobe, SAP, TIBCO o SUN, por citar algunas de ellas. Dentro del proyecto Enhydrase puede encontrar también Enhydra JaWE, un editor de workflow XPDL, según las especificaciones de WfMC.

Para obtener la funcionalidad de procesamiento analítico en línea (OLAP) se utilizan otras dos aplicaciones: el servidor OLAP Mondrian, que combinado con Jpivot, permiten realizar consultas a Data Marts, que los resultados sean presentados mediante un browser y que el usuario pueda realizar dril down y el resto de las navegaciones típicas.

Para el datamining, Pentaho incorpora la tecnología Weka la cual es una herramienta extensible e integrable que incluye herramientas para realizar transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización.

Otro de los módulos con que cuenta la Suite Pentaho es las ETL, el cual usa Kettle para realizar las transformaciones e integración de los datos. Esta herramienta es fácil de generar, mantener y desplegar. Cada proceso es creado con una herramienta gráfica a la que se le puede especificar qué hacer sin necesidad de escribir código para indicar cómo hacerlo. Es la herramienta, de código abierto, para integración de datos más utilizada en el mundo.

[2] Conjunto de protocolos y estándares que sirven para intercambiar datos entre aplicaciones.

1.12. Conclusiones parciales

En este capítulo se realizó un estudio teórico y conceptual sobre las herramientas y metodologías necesarias para la creación de un almacén de datos.

Sobre la base de este estudio, se decidió utilizar el enfoque de Raplh Kimball para el diseño del Data Mart, pues es menos costoso, más funcional, se pueden priorizar las áreas críticas; además su estructura de datos ofrece mayor facilidad al usuario para la exploración y búsqueda de información.

Se considera que la mejor y más completa arquitectura de datos es la de dos capas teniendo en cuenta que el almacén de datos será poblado desde un único origen de datos, con poco volumen de transformaciones. La organización física será ROLAP ya que con esta pueden hacerse un gran número de dimensiones y soporta elevados volúmenes de datos elementales, además de ser el motor OLAP de Pentaho.

Como herramienta de desarrollo para el Data Mart se seleccionó la Suite de Business Intelligence Pentaho en su versión 3.5 por ser de código abierto, ser actualmente la líder open source en temas de depósitos de datos y ser una plataforma orientada a la solución y centrada en procesos.

Teniendo en cuenta las características de los gestores de bases de datos anteriormente expuestos se utilizará Oracle 10g pues es capaz de manejar eficientemente todo el volumen de información del Data Mart y proporciona todo el soporte necesario para la construcción y mantenimiento del mismo, además de ser el gestor de base de datos utilizado en el proyecto donde en un futuro será implantado el Data Mart.

CAPÍTULO 2: DISEÑO DEL DATA MART

2. Introducción

En el presente capítulo se lleva a cabo el diseño del Data Mart para la Sala Situacional del SIGEP desde el punto de vista de Clasificación y Atención Integral. Se hace una descripción detallada del mismo y se profundiza en el proceso de diseño de los almacenes de datos utilizando la metodología DWEP.

2.1. Descripción del Data Mart

Se diseñó un Data Mart para la Clasificación y la Atención Integral de la sala situacional del SIGEP de manera que ayuda a la toma de decisiones a la dirección del sistema de justicia venezolano, enfocado a organizar y sistematizar la atención que recibirán los privados de libertad durante su permanencia en el sistema penitenciario. El mismo contiene información referente a los individuos, los centros penitenciarios, las unidades productivas que se conforman en el establecimiento penitenciario, las actividades educativas desarrolladas en el establecimiento, las disciplinas deportivas que se practican en el penal, la incorporación y participación de los privados de libertad en las manifestaciones culturales que se practican en el centro así como la clasificación de los penados en alguno de los regímenes de seguridad penal (máxima, media o mínima), para ofrecerle un atención individualizada.

El Data Mart sobre la Clasificación y la Atención Integral de la Sala Situacional del SIGEP, además de los reportes predefinidos, ofrece 5 variantes diferentes para la exploración de los datos. Los usuarios finales podrán explorar basándose en la clasificación, en este se muestran los individuos, sus expedientes y la clasificación de los individuos en cada uno de los centros penitenciarios modelados en el tiempo. De igual manera se puede explorar teniendo en cuenta la actividad cultural; esta exploración incluye individuos, sus expedientes, el nivel y el tipo de manifestación cultural practicada en el centro por el individuo a lo largo del tiempo. También se puede enfocar la exploración centrándose en la actividad productiva, incluyendo la información de

los individuos, sus expedientes, el tipo de actividad productiva que realiza en el centro a lo largo del tiempo de estancia en el penal. Otra manera de exploración es por la actividad deportiva, incluyendo individuos, sus expedientes, el nombre y tipo de disciplina deportiva en las que participa el individuo modelado en el tiempo. Por último los usuarios pueden explorar centrándose en la actividad educativa, la cual muestra a los individuos, sus expedientes y la clasificación de las actividades deportivas desarrolladas en el centro a través del tiempo.

Ver anexo 1 al 5

2.2. Proceso de diseño de un almacén de datos

Se puede dividir el proceso de diseño de un almacén de datos en 3 etapas: diseño conceptual, diseño lógico y diseño físico. En cuanto al tipo de consulta y el rendimiento esperado, los almacenes de datos y las bases de datos operacionales, difieren, lo que hace que las estrategias de diseño y el modelado de los datos sean diferentes.

“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 sistema de gestión de base de datos. Por último, en la etapa de diseño físico se implementa el esquema lógico en el manejador de bases de datos elegido.” (5)

2.3. Aplicación del método DWEP

El Data Warehouse Engineering Process es un método para llevar a cabo el diseño de un almacén de datos incluyendo las fuentes de datos operacionales, los procesos ETL y el propio esquema del almacén de datos. A continuación se muestran los flujos de trabajo del ciclo de vida del Data Mart para la Clasificación y Atención Integral de la Sala Situacional del SIGEP, así como los diagramas que los componen.

2.3.1. Requerimientos

En este flujo de trabajo se define el alcance del Data Warehouse, se recopilan los requisitos iniciales de los usuarios. Los requisitos se obtuvieron mediante el análisis de los reportes sobre la clasificación y la atención integral solicitados por los clientes en el flujo de trabajo Requerimiento en la fase de Inicio del proyecto SIGEP.

2.3.2. Análisis

El flujo de trabajo Análisis tiene como entrada los requisitos obtenidos en el flujo de Requerimientos. Sus objetivos son refinar y estructurar estos requerimientos y definir las fuentes de datos operacionales que sirven como fuente al Data Mart.

En la modelación de las fuentes de datos en diferentes niveles de detalles se realiza el esquema conceptual (SCS) y el esquema lógico (SLS). Para poblar el Data Mart se utiliza la base de datos operacional del SIGEP, donde se encuentra de forma normalizada toda la información requerida.

2.3.2.1. Esquema conceptual de la fuente (SCS)

Este esquema tiene como objetivo conocer qué datos están disponibles para el almacén de datos, para ello se representan las entidades más importantes y las relaciones entre ellas. (Ver anexo 6)

2.3.2.2. Esquema lógico de la fuente (SLS)

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

2.3.3. Diseño

En este flujo de trabajo se construye el modelo conceptual del Data Warehouse (DWCS). Para llevar a cabo esta actividad existen dos estrategias de procesamiento de información: *top-down y bottom-up.*

La primera de estas estrategias es la **top-down,** la cual define al almacén de datos según los requisitos de los usuarios finales y la otra es la **bottom-up**, en la cual se define el almacén de datos en base a los datos disponibles en las fuentes de datos. La estrategia que se siguió fue la *top down*, ya que se tuvieron en cuenta los requisitos funcionales obtenidos en el flujo de trabajo Requerimiento del proyecto SIGEP.

Luego de tener el esquema conceptual del almacén de datos, se define el esquema conceptual del cliente, que puede implementarse como Data Mart reales o virtuales.

Para el mapeo de datos entre las diferentes fuentes de datos (SCS) y el almacén de datos (DWCS), y el mapeo de este último con el esquema conceptual del cliente, se definen los procesos de extracción, transformación y carga o ETL.

2.3.3.1. Esquema conceptual del almacén de datos

Este esquema tiene como objetivo seleccionar los aspectos más importantes para la toma de decisiones y especificar cómo utilizarlos en dimensiones y/o medidas.

El método DWEP establece el proceso de diseño en tres niveles:

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 de 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 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 Data Mart está compuesto por 5 tablas de hechos: Tbl_Clasificación, Tbl_Actividad Educativa, Tbl_Actividad_Cultural, Tbl_Actividad_Productiva, Tbl_Actividad_Deportiva, y por trece dimensiones: Dim_Expediente, Dim_Centro, Dim_Individuo, Dim_Tiempo, Dim_Entrevista, Dim_Nombre_Clasificación, Dim_Tipo_AP, Dim_Clasificacion_AE, Dim_Tipo_AC, Dim_Nivel_AC, Dim_Manifestación_AC, Dim_Disciplina_AD y Dim_Tipo_AD.

En la figura 8 se muestra el esquema conceptual del almacén de datos a nivel 1, mostrando los diferentes diagramas de estrellas que forman el Data Mart: Shema_Clasificación, Schema_Actividad_Cultural, Schema_Actividad_Productiva, Schema_Actividad_Deportiva, Schema_Actividad_Educativa.

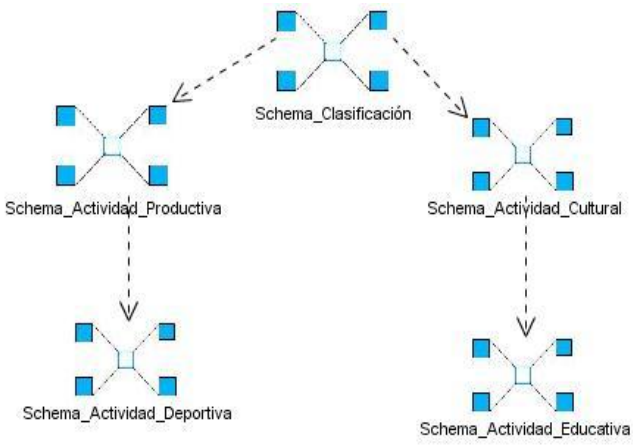


Figura 8: Esquema conceptual del Data Mart (Nivel 1)

En la figura 9 se representa el esquema conceptual del paquete FackPack_Clasificacion, compuesto por la tabla de hechos Tbl_Clasificacion y las dimensiones que esta utiliza: Dim_Individuo, Dim_Centro, Dim_Tiempo, Dim_Entrevista, Dim_Expediente, Dim_Nombre_Clasificacion.

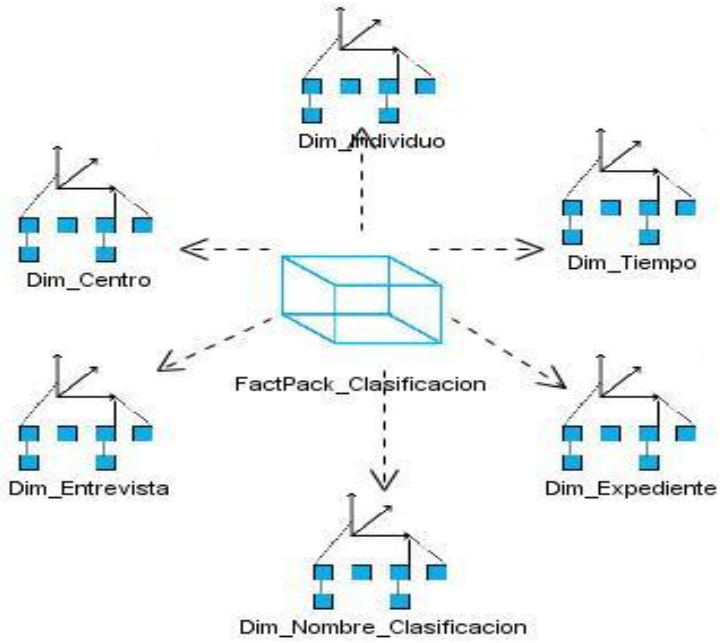


Figura 9: Esquema conceptual del Data Mart. Schema_Clasificacion (Nivel 2)

La figura 10 muestra el esquema conceptual del paquete FactPack_Actividad_Cultural compuesto por la tabla de hechos Tbl_Actividad_Cultural y las dimensiones que la componen: Dim_Individuo, Dim_Centro, Dim_Tiempo, Dim_Tipo_AC, Dim_Expediente, Dim_Nivel_AC.

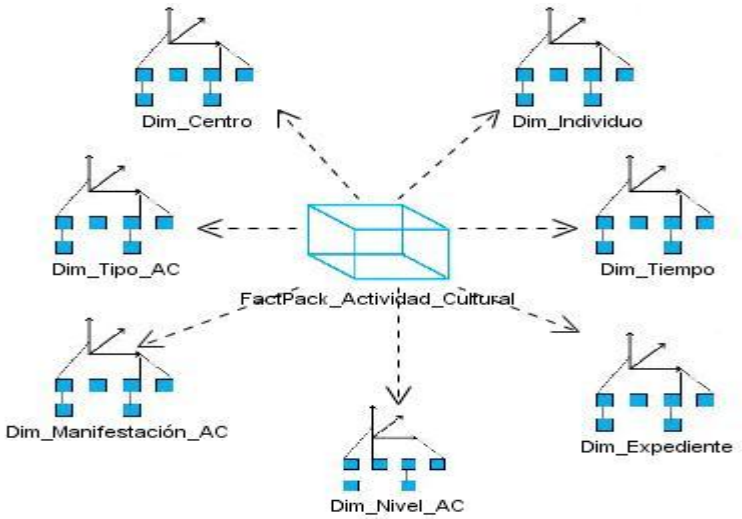


Figura 10: Esquema conceptual del Data Mart. Schema_Actividad_Cultural (Nivel 2)

En la figura 11 se representa el esquema conceptual del paquete FactPack_Actividad_Deportiva, compuesto por la tabla de hechos Tbl_Actividad_Deportiva y las dimensiones que esta utiliza: Dim_Individuo, Dim_Centro, Dim_Tiempo, Dim_Expediente, Dim_Tipo_AD, Dim_Discilplina_AD.

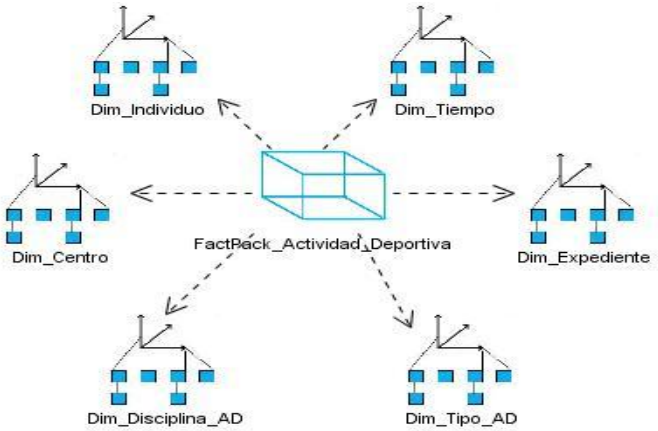


Figura 11: Esquema conceptual del Data Mart. Schema_Actividad_Deportiva (Nivel 2)

En la figura 12 se representa el esquema conceptual del paquete FactPack_Actividad_Educativa, compuesto por la tabla de hechos Tbl_Actividad_Educativa y las dimensiones que esta utiliza: Dim_Individuo, Dim_Tiempo, Dim_Expediente, Dim_Centro, Dim_Clasificacion_AE.

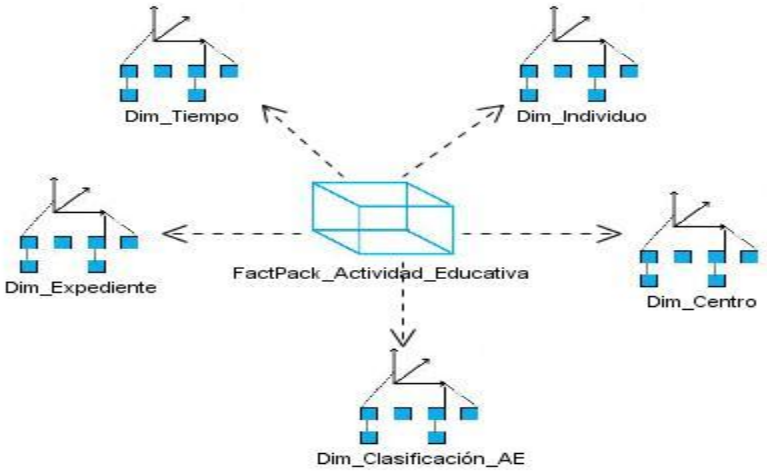


Figura 12: Esquema conceptual del Data Mart. Schema_Actividad_Educativa (Nivel 2)

En la figura 13 se representa el esquema conceptual del paquete FactPack_Actividad_Productiva, compuesto por la tabla de hechos Tbl_Actividad_Productiva y las dimensiones que esta utiliza: Dim_Individuo, Dim_Tiempo, Dim_Expediente, Dim_Centro, Dim_Tipo_AP.

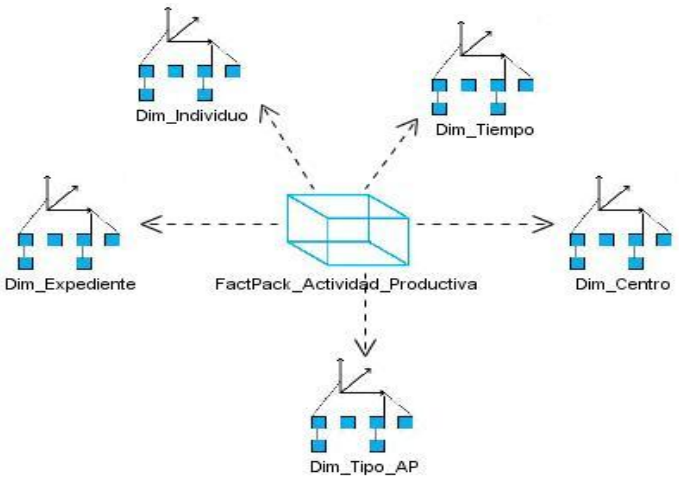


Figura 13: Esquema conceptual del Data Mart. Schema_Actividad_Productiva (Nivel 2)

El anexo 7 refleja el contenido del paquete FactPack_Clasificacion a nivel 3, que no es más que la tabla de hecho Tbl_Clasificacion con sus correspondientes medidas (cantidad) y las dimensiones que utiliza.

El anexo 8 refleja el contenido del paquete FactPack_Actividad_Cultural a nivel 3, que no es más que la tabla de hecho Tbl_Actividad_Cultural con sus correspondientes medidas (cantidad) y las dimensiones que utiliza.

El anexo 9 refleja el contenido del paquete FactPack_Actividad_Deportiva a nivel 3, que no es más que la tabla de hecho Tbl_Actividad_Deportiva con sus correspondientes medidas (cantidad) y las dimensiones que utiliza.

El anexo 10 refleja el contenido del paquete FactPack_Actividad_Educativa a nivel 3, que no es más que la tabla de hecho Tbl_Actividad_Educativa con sus correspondientes medidas (cantidad) y las dimensiones que utiliza.

El anexo 11 refleja el contenido del paquete FactPack_Actividad_Productiva a nivel 3, que no es más que la tabla de hecho Tbl_Actividad_Productiva con sus correspondientes medidas (cantidad) y las dimensiones que utiliza.

En el anexo 12 se muestra una vista global de todas las definiciones de los esquemas estrellas en un solo diagrama. Se encuentran las tablas de hechos y las dimensiones que contienen el Data Mart así como las dependencias entre estas.

2.4. Mapeo de datos

"El diagrama de mapeo de datos (Data Mapping) es un nuevo tipo de diagrama adaptado para representar el flujo de datos, con varios niveles de detalle en un almacén de datos. Para capturar las interconexiones entre los elementos del diseño, en término de los datos, usamos la noción de *mapeo*." (10)

Como un diagrama de mapeo de datos puede tornarse muy complejo, Lujan-Mora (10) lo divide en cuatro niveles:

- **Nivel 0 ó de base de datos**: 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 1 ó de flujo de datos**: 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 2 ó de tabla**: 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 3 ó de atributo**: En este nivel, el diagrama de mapeo de datos captura los mapeos existentes a nivel de atributo.
- En la figura14se encuentra representado el mapeo de datos a nivel de base de datos o Nivel 0. Este está representado por un paquete nombrado Data Mapping, donde están todas las operaciones de mapeo, otro denominado DWCS que constituye el modelo conceptual del Data Warehouse y un tercer paquete llamado SCS que representa el esquema conceptual de la fuente.

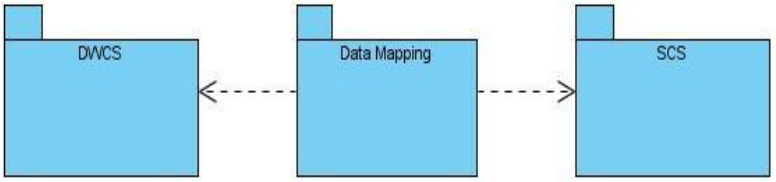


Figura 14: Mapeo de datos a nivel de base de datos (Nivel 0)

El anexo 13 muestra a nivel de tablas cómo se puebla la Dim_Centro. Se utilizan 3 tablas de la fuente trasladando los datos de las mismas para la tabla Dim_Centro.

El anexo 14 muestra a nivel de tablas cómo se puebla la Dim_Expediente. Se utiliza 1 tabla de la fuente trasladando los datos de la misma para la tabla Dim_Centro.

El anexo 15 muestra a nivel de tablas cómo se puebla la Dim_Tiempo con las fechas en que se realizan las actividades educativas necesarias para poblar la tabla Tbl_Actividad_Educativa.

El anexo 16 muestra a nivel de tablas cómo se puebla la Dim_Individuo. Se utilizan 3 tablas de la fuente trasladando los datos de las mismas para la tabla Dim_Individuo.

El anexo 17 muestra a nivel de tablas cómo se puebla la Dim_Tiempo con las fechas en que se realizan las clasificaciones necesarias para poblar la tabla Tbl_Clasificación.

El anexo 18 muestra a nivel de tablas cómo se puebla la Dim_Tiempo con las fechas en que se realizan las actividades culturales necesarias para poblar la tabla Tbl_Actividad_Cultural.

El anexo 19 muestra a nivel de tablas cómo se puebla la Dim_Tiempo con las fechas en que se realizan las actividades deportivas necesarias para poblar la tabla Tbl_Actividad_Deportiva.

El anexo 20 muestra a nivel de tablas cómo se puebla la Dim_Tiempo con las fechas en que se realizan las actividades productivas necesarias para poblar la tabla Tbl_Actividad_Productiva.

El anexo 21 muestra a nivel de tablas cómo se puebla la Dim_Clasificacion_AE. Se utilizan 6 tablas de la fuente trasladando los datos de las mismas para la tabla Dim_Clasificacion_AE.

El anexo 22 muestra a nivel de tablas cómo se puebla la Dim_Entrevista. Se utilizan 2 tablas de la fuente trasladando los datos de las mismas para la tabla Dim_Entrevista.

El anexo 23 muestra a nivel de tablas cómo se puebla la Dim_Manifestación_AC. Se utiliza 1 tabla de la fuente trasladando los datos de la misma para la tabla Dim_Manifestación_AC.

El anexo 24 muestra a nivel de tablas cómo se puebla la Dim_Disciplina_AD. Se utiliza 1 tabla de la fuente trasladando los datos de la misma para la tabla Dim_Disciplina_AD.

El anexo 25 muestra a nivel de tablas cómo se puebla la Dim_Clasificacion. Se utiliza 1 tablas de la fuente trasladando los datos de la misma para la tabla Dim_Clasificacion.

El anexo 26 muestra a nivel de tablas cómo se puebla la Dim_Tipo_AC. Se utiliza 1 tabla de la fuente trasladando los datos de la misma para la tabla Dim_Tipo_AC.

El anexo 27 muestra a nivel de tablas cómo se puebla la Dim_Tipo_AD. Se utiliza 1 tabla de la fuente trasladando los datos de la misma para la tabla Dim_Tipo_AD.

El anexo 28 muestra a nivel de tablas cómo se puebla la Dim_Tipo_AP. Se utiliza 1 tabla de la fuente trasladando los datos de la misma para la tabla Dim_Tipo_AP.

El anexo 29 representa a nivel de tablas cómo se puebla la tabla Tbl_Actividad_Cultural, para ello se extraen datos de las dimensiones Dim_Expediente, Dim_Tipo_AC, Dim_Manifestacion_AC, Dim_Individuo, Dim_Tiempo, Dim_Centro, Dim_Nivel_AC.

El anexo 30 representa a nivel de tablas cómo se puebla la tabla Tbl_Actividad_Deportiva, para ello se extraen datos de las dimensiones Dim_Expediente, Dim_Tipo_AD, Dim_Disciplina_AD, Dim_Individuo, Dim_Tiempo, Dim_Centro.

El anexo 31 representa a nivel de tablas cómo se puebla la tabla Tbl_Actividad_Educativa, para ello se extraen datos de las dimensiones Dim_Expediente, Dim_Clasificacion_AE, Dim_Individuo, Dim_Tiempo, Dim_Centro.

El anexo 32 representa a nivel de tablas cómo se puebla la tabla Tbl_Actividad_Productiva, para ello se extraen datos de las dimensiones Dim_Expediente, Dim_Tipo_AP, Dim_Individuo, Dim_Tiempo, Dim_Centro.

El anexo 33 representa a nivel de tablas cómo se puebla la tabla Tbl_Clasificacion, para ello se extraen datos de las dimensiones Dim_Expediente, Dim_Entrevista, Dim_Individuo, Dim_Tiempo, Dim_Centro, Dim_Nombre_Clasificacion.

2.5. Proceso de Extracción, Transformación y Carga.

Como parte del flujo de trabajo Implementación se lleva a cabo la construcción de las estructuras físicas del Data Warehouse, definiendo el tipo de almacenamiento empleado así como creando los procesos ETL necesarios para poblar el Data Warehouse.

2.5.1. Construcción del sistema de almacenamiento de datos

Se monta el Gestor de Base de Datos Oracle 10g de acuerdo al diseño propuesto, por lo que se crean las tablas con sus respectivos atributos y relaciones de manera que se define un sistema de almacenamiento adecuado para los procesos que se sucederán. Dicha base de datos

multidimensional contendrá la información referente a la Clasificación y Atención Integral de la Sala Situacional del SIGEP, la cual sería actualizada de forma diaria.

2.5.2. Carga de los datos en el almacén de datos

“El proceso de población de un almacén de datos requiere la utilización de los procesos ETL. La primera parte de este proceso consiste en extraer los datos desde los sistemas de origen, generalmente bases de datos relacionales, pero pueden provenir de no relacionales u otras estructuras diferentes, por esta razón es obligatorio verificar que los datos a extraer cumplan con las pautas que se esperaban. Además de estos chequeos de estructura, es necesario que las tareas de extracción no colapsen al sistema de origen, provocando que este no pueda utilizarse en su uso cotidiano.” (5)

La segunda fase de estos procesos es transformar, que no es más que aplicar algunas funciones o reglas a los datos extraídos para que estos puedan ser cargados en el almacén de datos. Estas transformaciones tienen como objetivo general: la corrección de errores, eliminación de redundancia y resolución de inconsistencias. Algunas de las transformaciones que se aplican con este fin son:

- Seleccionar sólo ciertas columnas para su carga (por ejemplo, que las columnas con valores nulos no se carguen).
- Traducir códigos (por ejemplo, si la fuente almacena una "H" para Hombre y "M" para Mujer pero el destino tiene que guardar "1" para Hombre y "2" para Mujer)
- Obtener nuevos valores calculados (por ejemplo, total_venta = cantidad * precio).
- Unir datos de múltiples fuentes (por ejemplo, búsquedas, combinaciones, etc.).
- Dividir una columna en varias (por ejemplo, columna "Nombre: García, Miguel"; pasar a dos columnas "Nombre: Miguel" y "Apellido: García").

Por último, una vez que los datos son extraídos y transformados sólo queda cargarlos en el almacén de datos. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes. En algunas bases de datos se sobrescribe la

información antigua con nuevos datos, en otras se crea un historial con la información para disponer de un rastro de la misma a través del tiempo.

La fase de carga interactúa directamente con la base de datos de destino. Al realizar esta operación se aplicarán todas las restricciones y triggers (disparadores) que se hayan definido en ésta (por ejemplo, valores únicos, integridad referencial, campos obligatorios, rangos de valores). Estas restricciones y triggers (si están bien definidos) contribuyen a que se garantice la calidad de los datos en el proceso ETL, y deben ser tenidos en cuenta.

En nuestro caso estos procesos ETL se implementarán haciendo uso de la herramienta de la Suite Pentaho, Kettle.

Para cada dimensión se implementó una transformación las que sirven para pasar los datos de la base de datos operacional hacia el almacén de datos.

En la figura 15 se muestra la transformación realizada a la dimensión Dim_Individuo.

Figura 15: Transformación Dim_Individuo

Esta transformación comienza utilizando el paso de Kettle **Entrada Tabla** (ver anexo 34) permitiendo especificar las tablas y los campos origen. Una vez definido esto, se procede a fusionar los nombres en un solo campo, este se realiza con el paso **Valor de Java Script Modificado** (ver anexo 35). Luego de tener bien definida la entrada de los datos se hacen una serie de búsquedas de varios valores en otras tablas mediantes el paso **Búsqueda en Base de Datos** (ver anexos 36 y 37). Los identificadores de las dimensiones son valores generados por una secuencia que se encuentra en la base de datos auxiliar, para obtener este valor se utiliza el paso **Añadir Secuencia** (ver anexo38). Al tener todos los datos de la dimensión se procede a mapearlos, utilizando el paso **Selecciona/Renombra Valores** (ver anexo 2.39), con los del paso **Salida Tabla** (ver anexo 40) para así terminar la transformación.

Las demás trasformaciones para poblar las otras dimensiones se implementaron de forma similar utilizando la mayoría de los pasos ya mencionados.

De manera general la parte más complicada de los procesos ETL es poblar las tablas de hechos. Para esto se realizaron 5 transformaciones. (Ver anexo 41 al 45).

Finalizadas todas las trasformación se ejecutan las mismas, esto se realiza dirigido por el trabajo ETL_Job (ver anexo 46). Este componente iniciará las transformaciones diariamente, a las 4:00am, con el fin que se estén utilizando las bases de datos operacionales lo menos posible. Para realizar la carga se llenan las dimensiones estándar en orden jerárquico, después la dimensión tiempo y por último las tablas de hechos.

2.6. Conclusiones parciales

En este capítulo se describió de manera detallada el Data Mart de la Sala Situacional del SIGEP orientado a la clasificación y atención integral, planteándose cómo se llevó a cabo el diseño del Data Mart mediante el uso de distintos diagramas propuestos en la metodología DWEPasí como la implementación de los procesos ETL necesarios para poblar el Data Mart, haciendo uso de la herramienta de la Suite Pentaho, Kettle.

CONCLUSIONES

- Un Data Warehouse da lugar a una serie importante de beneficios para la organización. Su utilización permite que la información de gestión sea: accesible, correcta, uniforme y actualizada.
- Se han cumplido en tiempo los objetivos trazados llevando a cabo de manera satisfactoria el diseño del Data Mart para la sala situacional del SIGEP desde el punto de vista de Clasificación y Atención Integral así como implementando los procesos ETL necesarios para poblar el almacén de datos.
- El modelo dimensional brinda una forma muy sencilla de representación de los datos y mejora así el tiempo de consulta a la base de datos.
- Mediante el diseño y posterior implementación de este Data Martse le dará soporte al proceso de toma de decisiones por parte del sistema penitenciario venezolano, al brindarle una herramienta de Inteligencia de Negocio eficiente y dinámica.
- El que se trate el desarrollo del Data Mart como un proceso iterativo e incremental permitirá que en un futuro se puedan incluir nuevas necesidades, procesos de negocios, así como cambios que se produzcan en el sistema penitenciario de Venezuela.

RECOMENDACIONES

- Continuar con la implementación de los cubos de datos asociados al diseño del Data Mart propuesto.

- Construir Data Marts a todos los módulos del proyecto para luego integrarlos en un Data Warehouse central que ayude al proceso de gestión de información y una acertada toma de decisiones.

- Extender la idea de construir Data Marts y Data Warehouse a todos los proyectos en desarrollo en la universidad.

- Motivar el estudio y desarrollo de los almacenes de datos que tanto auge han venido cobrando en los últimos tiempos.

REFERENCIAS BIBLIOGRÁFICAS

1. **Cano, Josep Lluís.** *Business Intelligence: Competir con información.* s.l. : Fundación Cultural Banesto, 2007.

2. **Arias, Arturo C.** *Visión del Sistema de Gestión Penitenciaria. Actualización.* La Habana : Universidad de las Ciencias Informáticas, 2007.

3. **Kimball, Ralph.** *The DW Lifecycle Toolkit, Second Edition.* s.l. : Wiley Computer Publishing, 2002.

4. **Inmon, William Harvey.** *Building the Data Warehouse, Fourth Edition.* s.l. : Wiley Publishing, Inc., 2005.

5. **Lazo Pedraja, David y Martínez Bravet, Rene.** *Repositorio de Datos de la Sala Situacional del SIGEP, desde el punto de vista del Control Penal.* [Digital] La Habana : s.n., 2010.

6. **Kimball, Ralph y Caserta, Joe.** *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* s.l. : Wiley Publishing, Inc., 2004.

7. **Ponniah, Paulraj.** *Data Warehousing Fundamentals for IT Professionals. Second Edition.* s.l. : John Wiley & Sons, Inc., 2010.

8. **Sinnexus.** Sinnexus. Business Intelligence + Informática Estratégica. [En línea] [Citado el: 21 de Octubre de 2010.] http://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx.

9. **Bernabeu, Ricardo Dario.** DATAPRIX. Knowledge Is The Goal. [En línea] 7 de Mayo de 2009. [Citado el: 3 de Marzo de 2011.] http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/-metodologia-hefesto/52-descripcion.

10. **Luján-Mora, Sergio.** *Diseño de Almacenes de Datos con UML.* Alicante : s.n., 2005.

11. **Jacobson, Ivar, Booch, Grady y Rumbaugh, James.** *El Proceso Unificado de Desarrollo de Software.* s.l. : Addison - Wesley, 2000.

12. Portada sobre plataforma Pentaho Open Source Business Intelligence. [En línea] 2008. [Citado el: 22 de Octubre de 2010.] http://pentaho.almacen-datos.com.

BIBLIOGRAFÍA

1. **Micosoft.***Guía Estratégica de Business Intelligence.* [Documento] 2004.
2. Business Intelligence Fácil. [En línea] 23 de Marzo de 2009. [Citado el: 22 de Octubre de 2010.] http://www.businessintelligence.info/definiciones/business-intelligence-system-1958.html.
3. **Wolff, Gloria Carmen.** La Tecnología Data Warehousing. [En línea] 2009. [Citado el: 18 de 11 de 2010.] http://www.inf.udec.cl/~revista/ediciones/edicion3/cwolff.PDF.
4. **Herrera, Basurto y Kirs, Cristhian.***Apuntes de Data Warehouse.* 2007.
5. **Iznaga González, Yonelbys.** Diseño e Implementación de un Data Warehouse para el Sistema de Gestión Estadística en Cuba. [En línea] 2008. [Citado el: 20 de Enero de 2011.] http://www.bibliodoc.uci.cu/TD/TD_1338_08.pdf.
6. **Bellatreche, Ladjel.***Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction.* New York : Information Science Reference, 2010. 978-1-60566-756-0.
7. **Hammergren, Thomas C. y Simon, Alan R.***Data Warehousing For Dummies. Second Edition.* Indianapolis : Wiley Publishing, Inc., 2009. 978-0-470-40747-9.
8. **Hobbs, Lilian, y otros, y otros.***Oracle Database 10g Data Warehousing.* s.l. : Elsevier Digital Press, 2005. 1-55558-322-9.
9. **Reeves, Laura L.***A Manager's Guide to Data Warehousing.* Indianapolis : Wiley Publishing, Inc, 2009. 978-0-470-17638-2.

10. **Wang, John.***Encyclopedia of Data Warehousing and Mining. Second Edition.* New York : Information Science Reference, 2009. 978-1-60566-010-3.

11. **Espinosa, Roberto.** El Rincón del BI. Descubriendo el Business Intelligence. [En línea] 5 de Diciembre de 2009. [Citado el: 20 de Octubre de 2010.] http://churriwifi.wordpress.com/2009/12/05/5-fases-en-la-implantacion-de-un-sistema-dw-metodologia-para-la-construccion-de-un-dw/.

12. **Díaz Morales, Themis Patricia y Bermúdez Rodríguez, José Salvador.** Diseño de un Data Warehouse en los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular . [En línea] 2010. [Citado el: 7 de Diciembre de 2010.] http://bibliodoc.uci.cu/TD/TD_02946_10.pdf.

13. **Waite, Anthony.** *Analytic Workspace Manager and Oracle OLAP 10g.* [Documento] s.l. : Oracle Corporation, 2004.

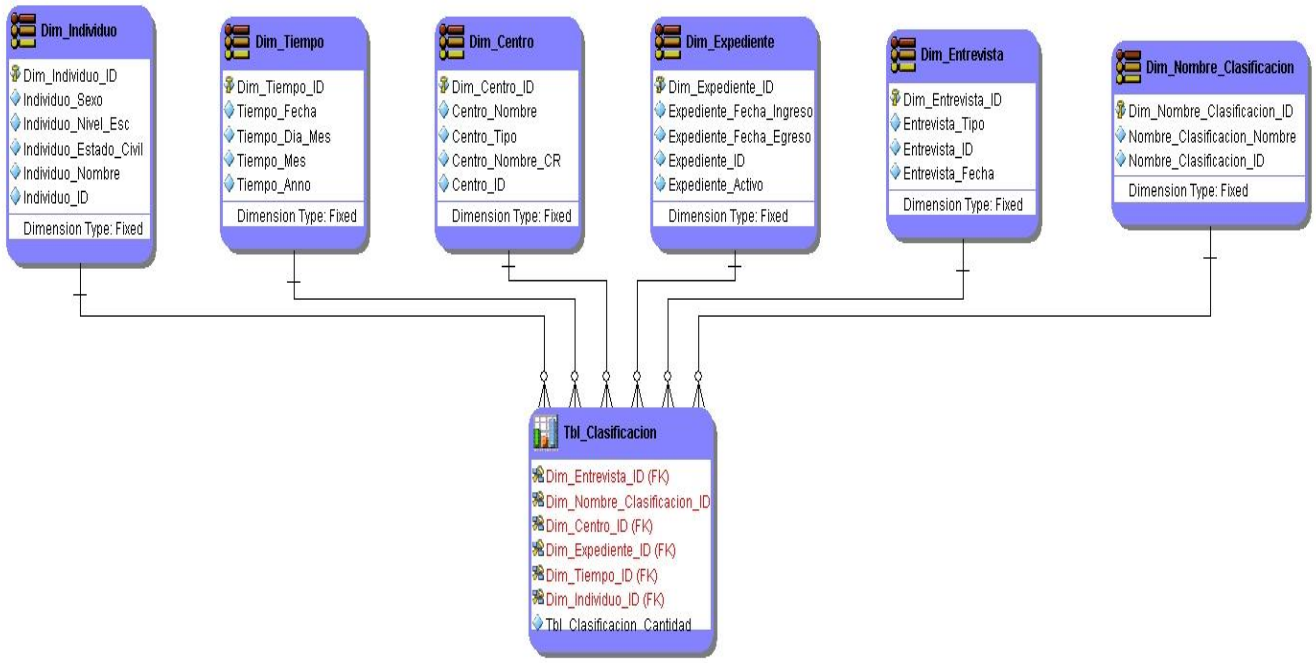
14. **Adamson, Christopher.** *Mastering Data Warehouse Aggregates Solutions for Star Schema Performance.* Indianapolis : Wiley Publishing, Inc., 2006. 978-0-471-77709-0.

15. **Wrembel, Robert y Koncilia, Christian.** *Data warehouses and OLAP : concepts, architectures, and solutions.* s.l. : Idea Group Inc, 2007. 1-59904-366-1.

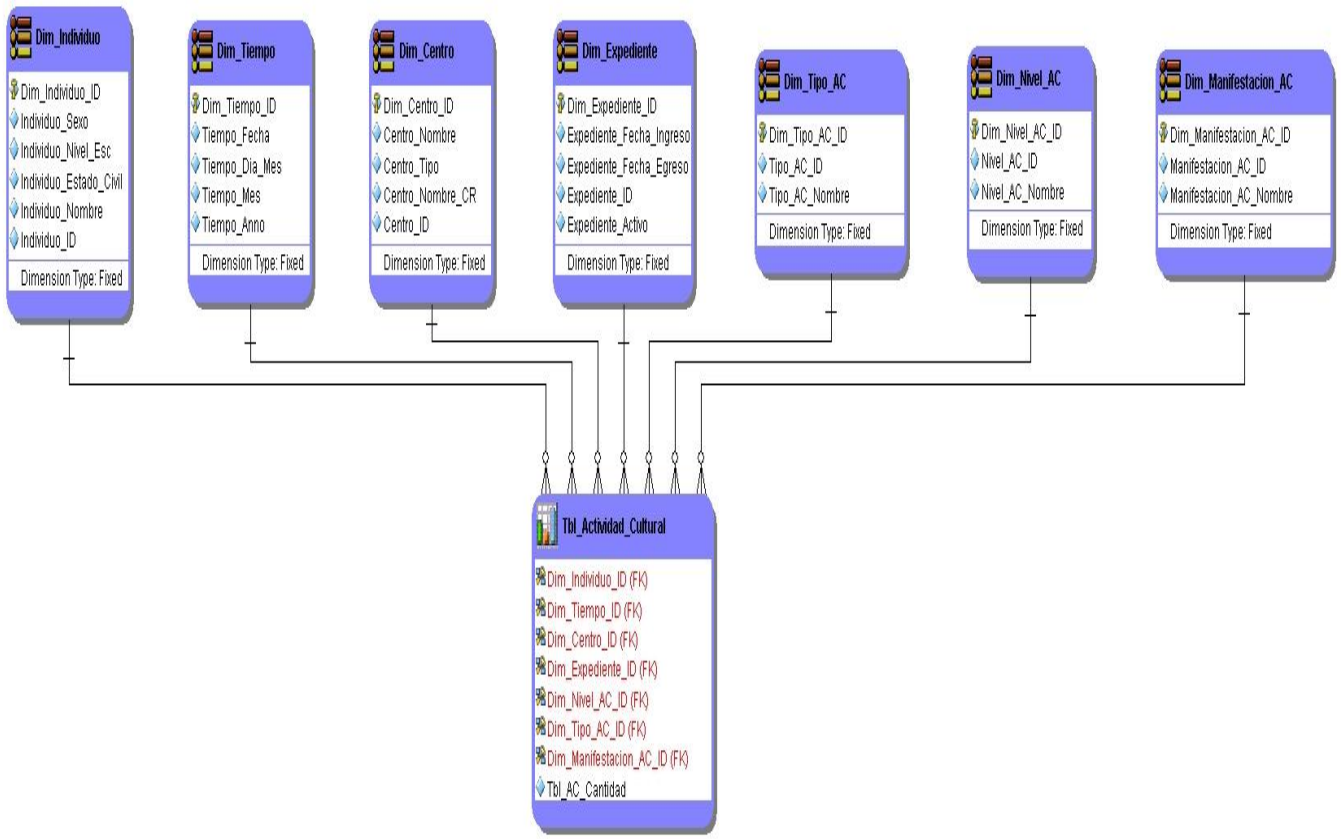
16. **Neera, Bhansali.** *STRATEGIC DATA WAREHOUSING: Achieving Alignment with Business.* s.l. : Auerbach Publications , 2010. 978-1-4200-8394-1.

ANEXOS

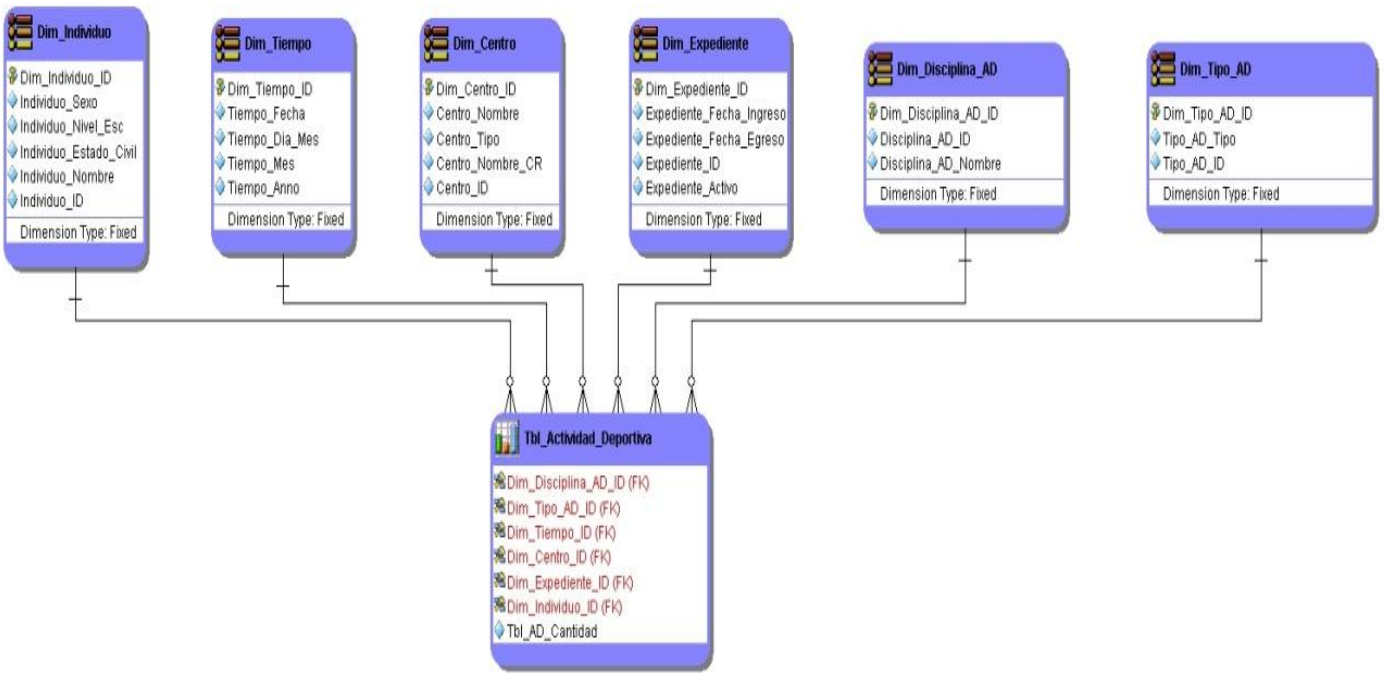
Anexo 1. Esquema tabla de hechos Clasificación



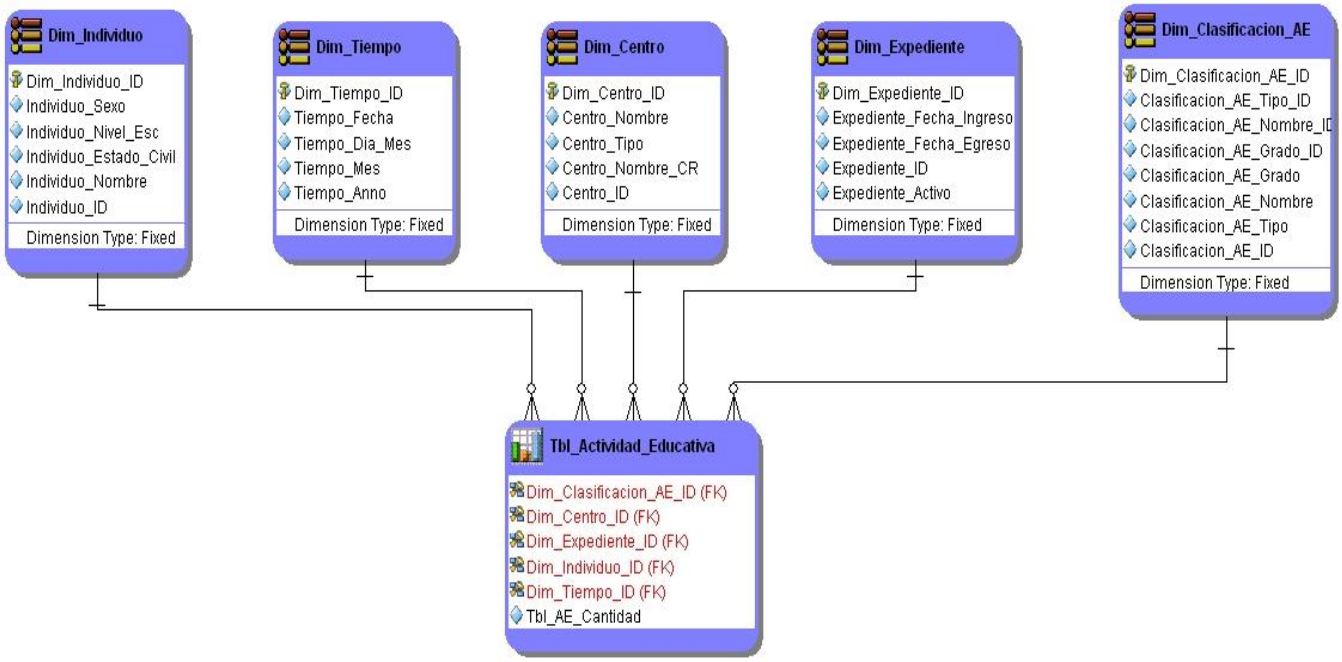
Anexo 2 Esquema tabla de hechos Actividad_Cultural



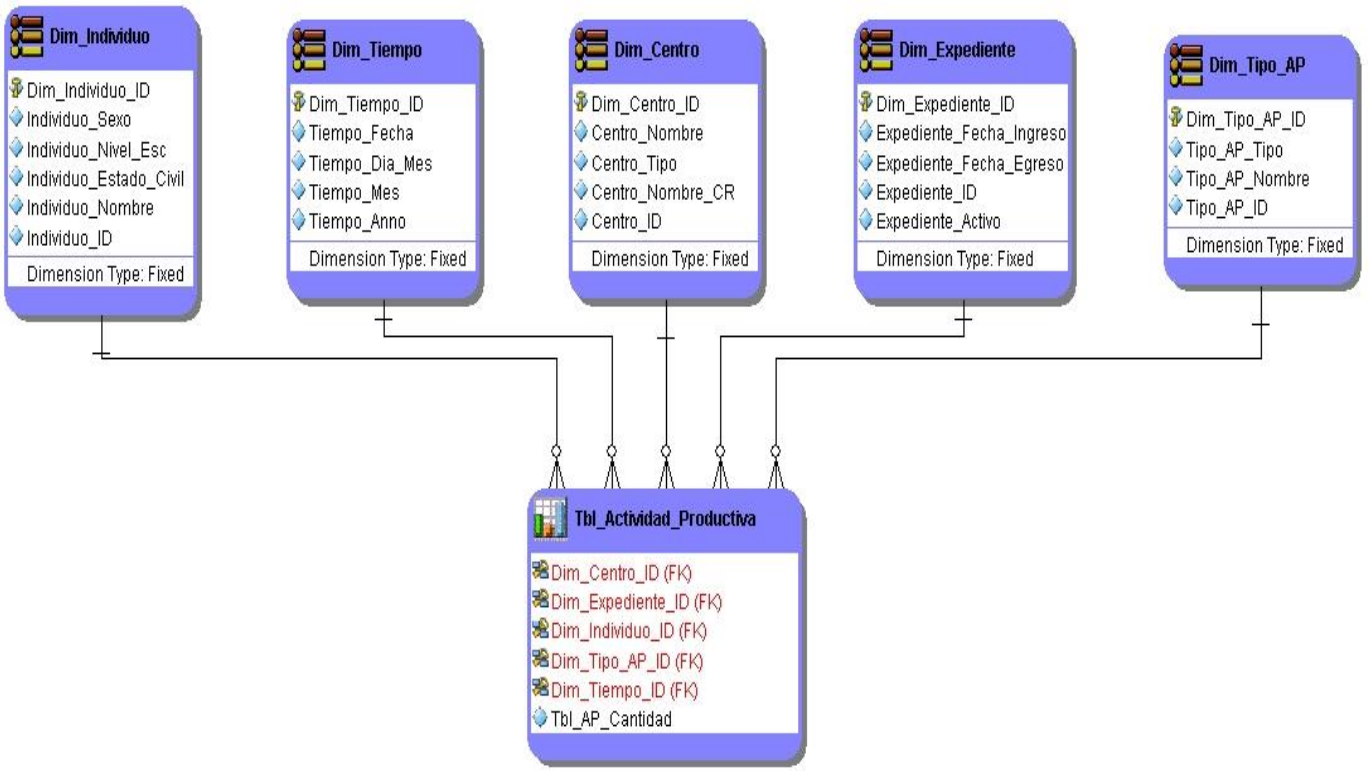
Anexo 3 Esquema tabla de hechos Actividad_Deportiva



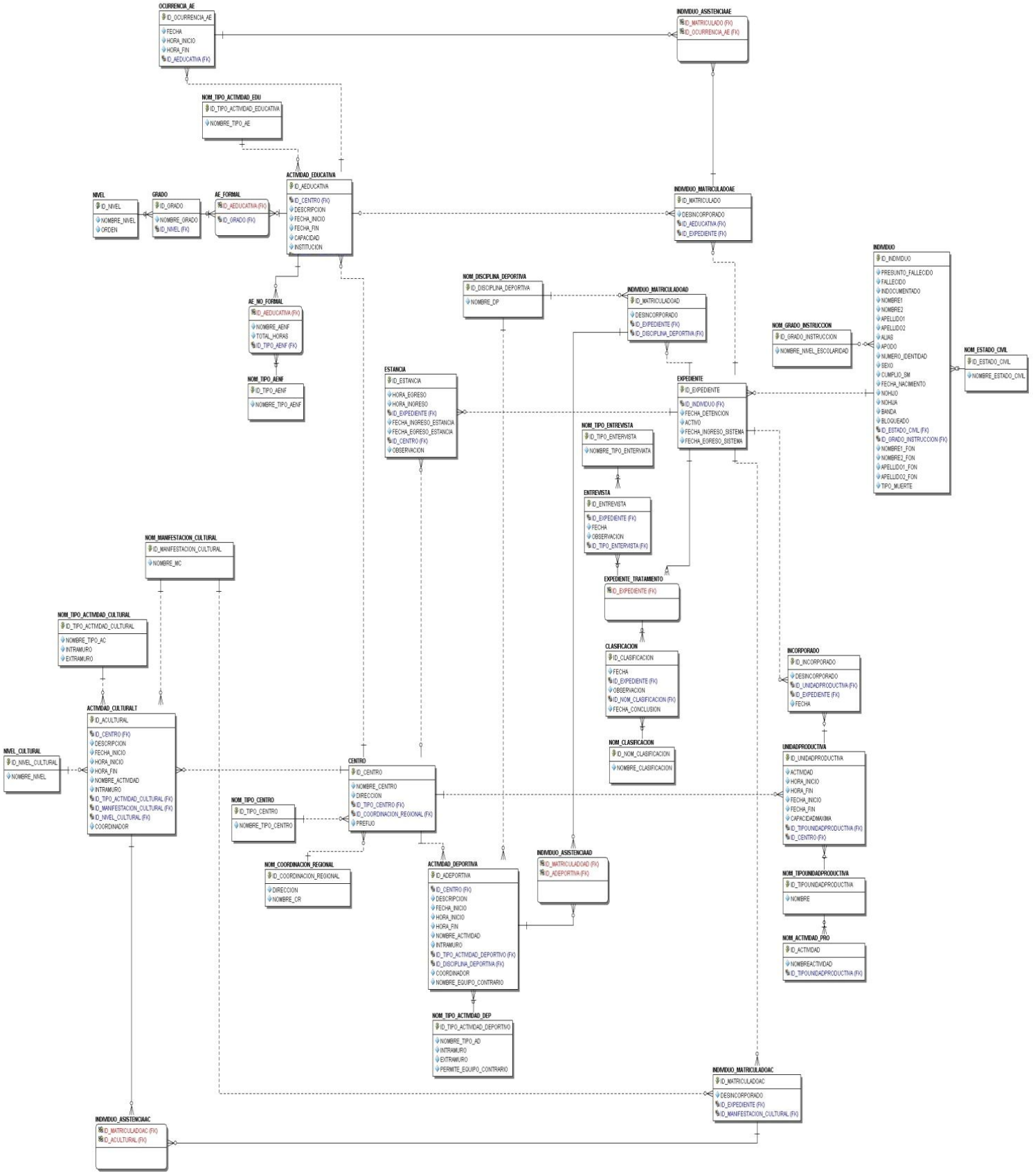
Anexo 4 Esquema tabla de hechos Actividad_Educativa



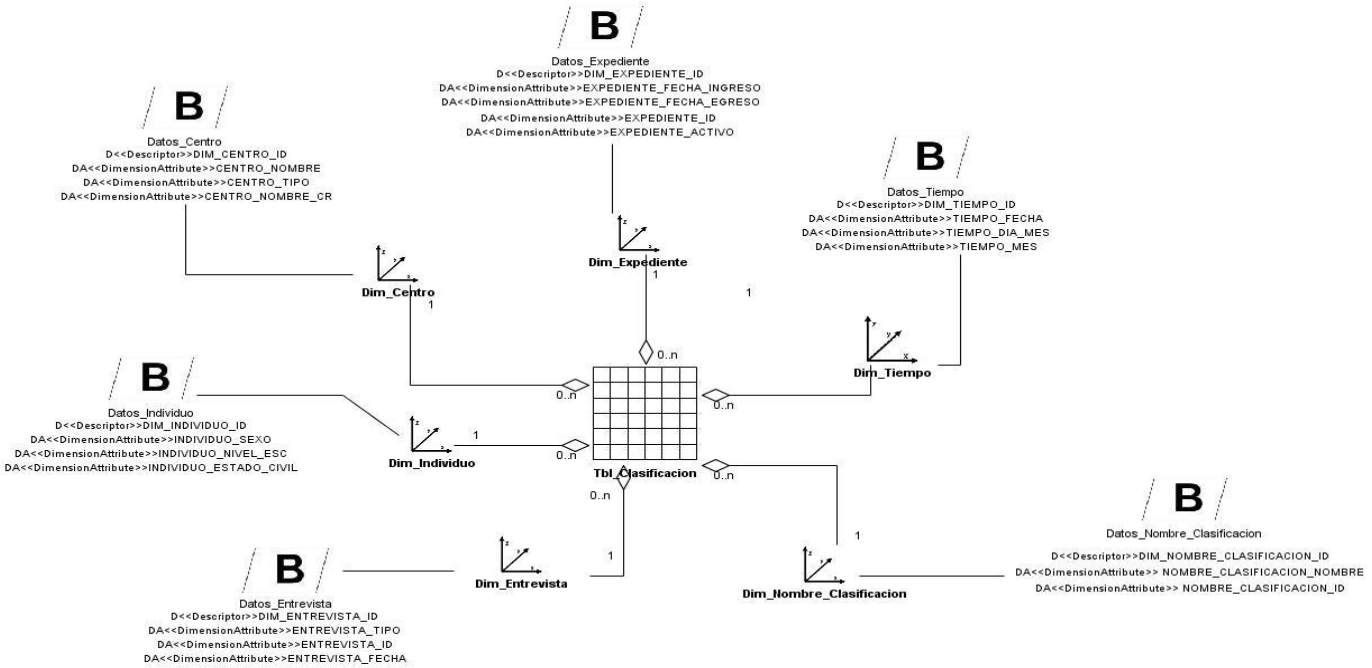
Anexo 5 Esquema tabla de hechos Actividad_Productiva



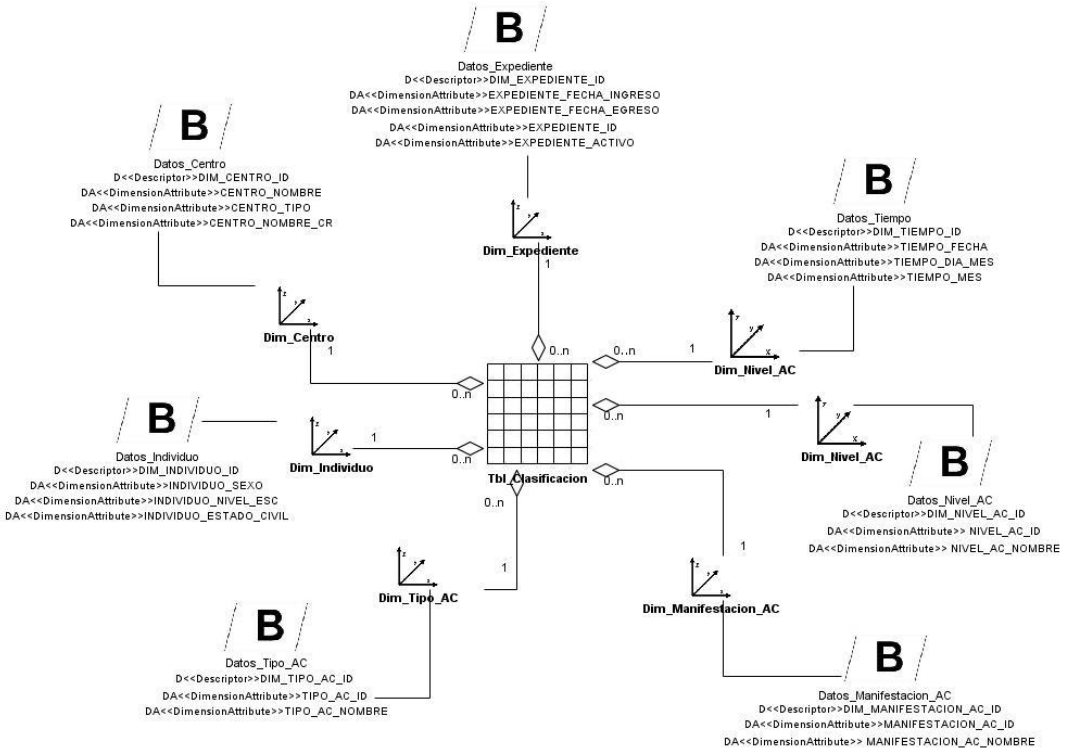
Anexo 6. Esquema conceptual de la fuente



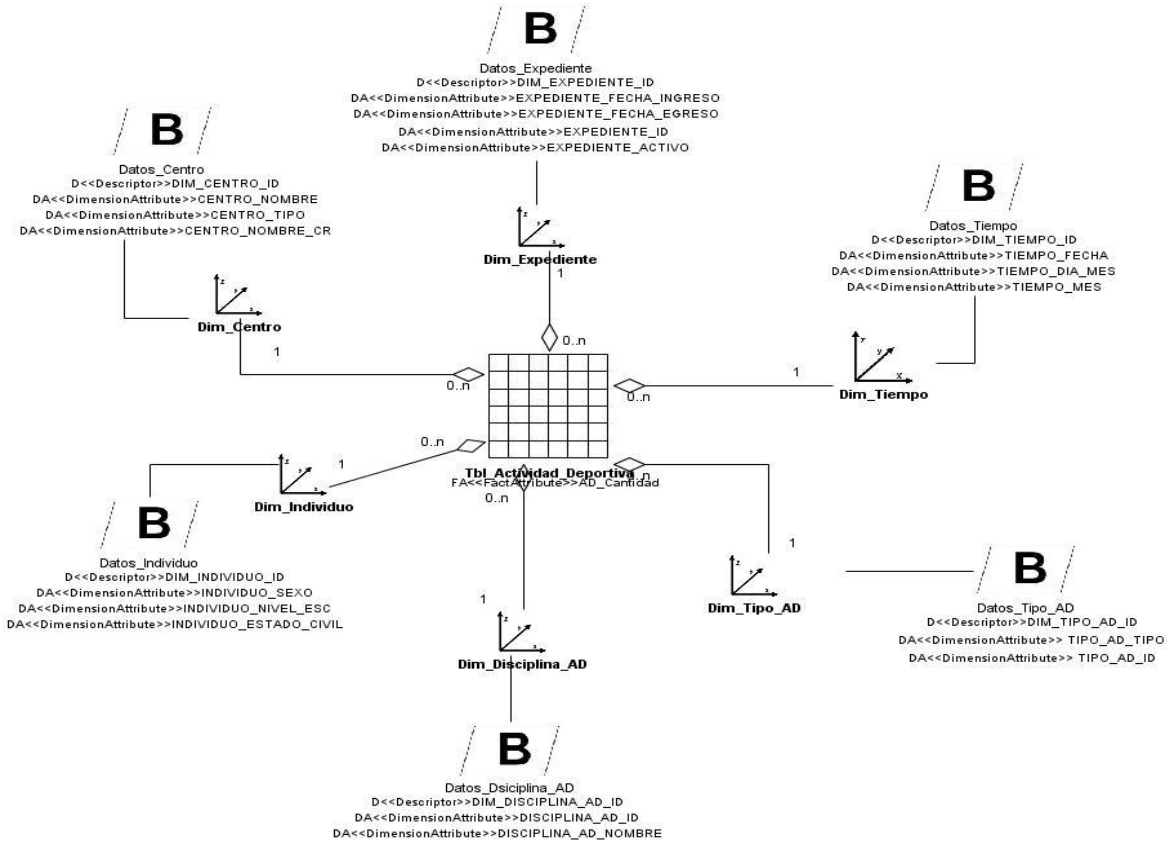
Anexo 7. Esquema conceptual del almacén de datos. Schema_Clasificacion (Nivel3).



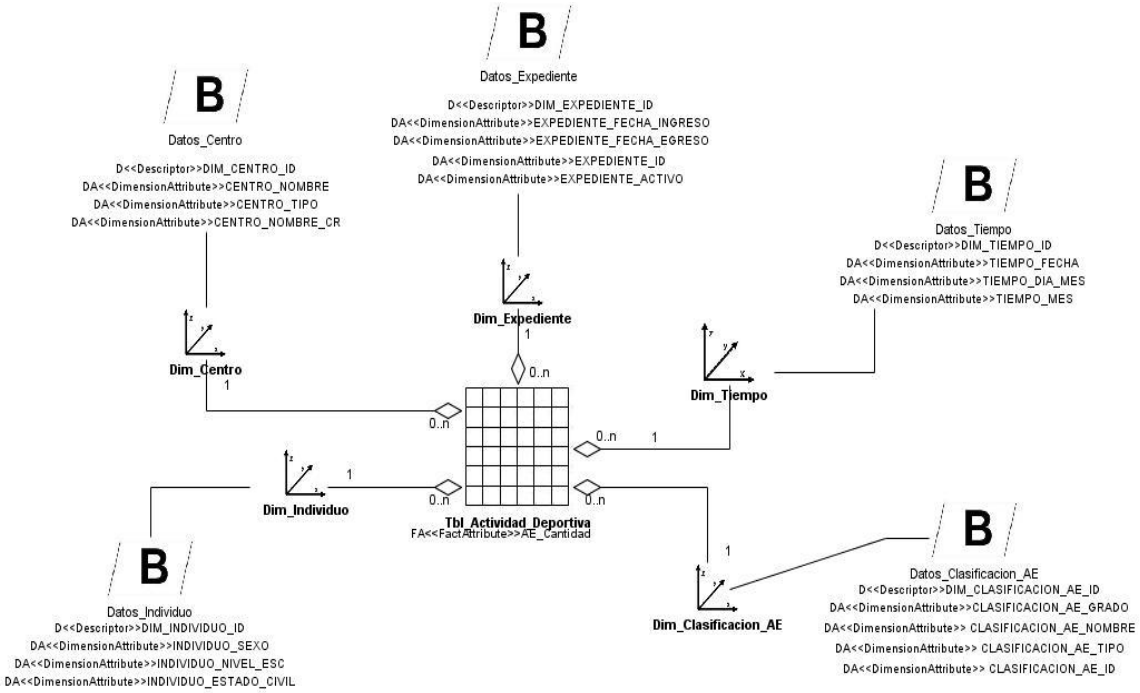
Anexo 8. Esquema conceptual del almacén de datos. Schema_Actividad_Cultural (Nivel 3)



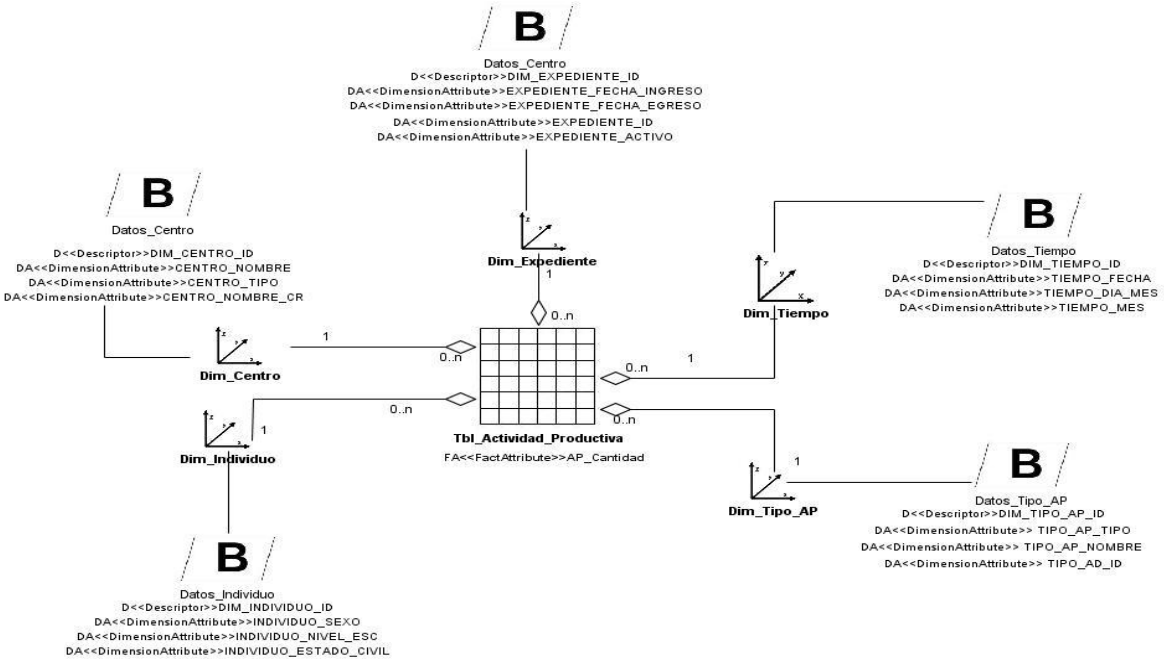
Anexo 9. Esquema conceptual del almacén de datos. Schema_Actividad_Deportiva (Nivel 3)



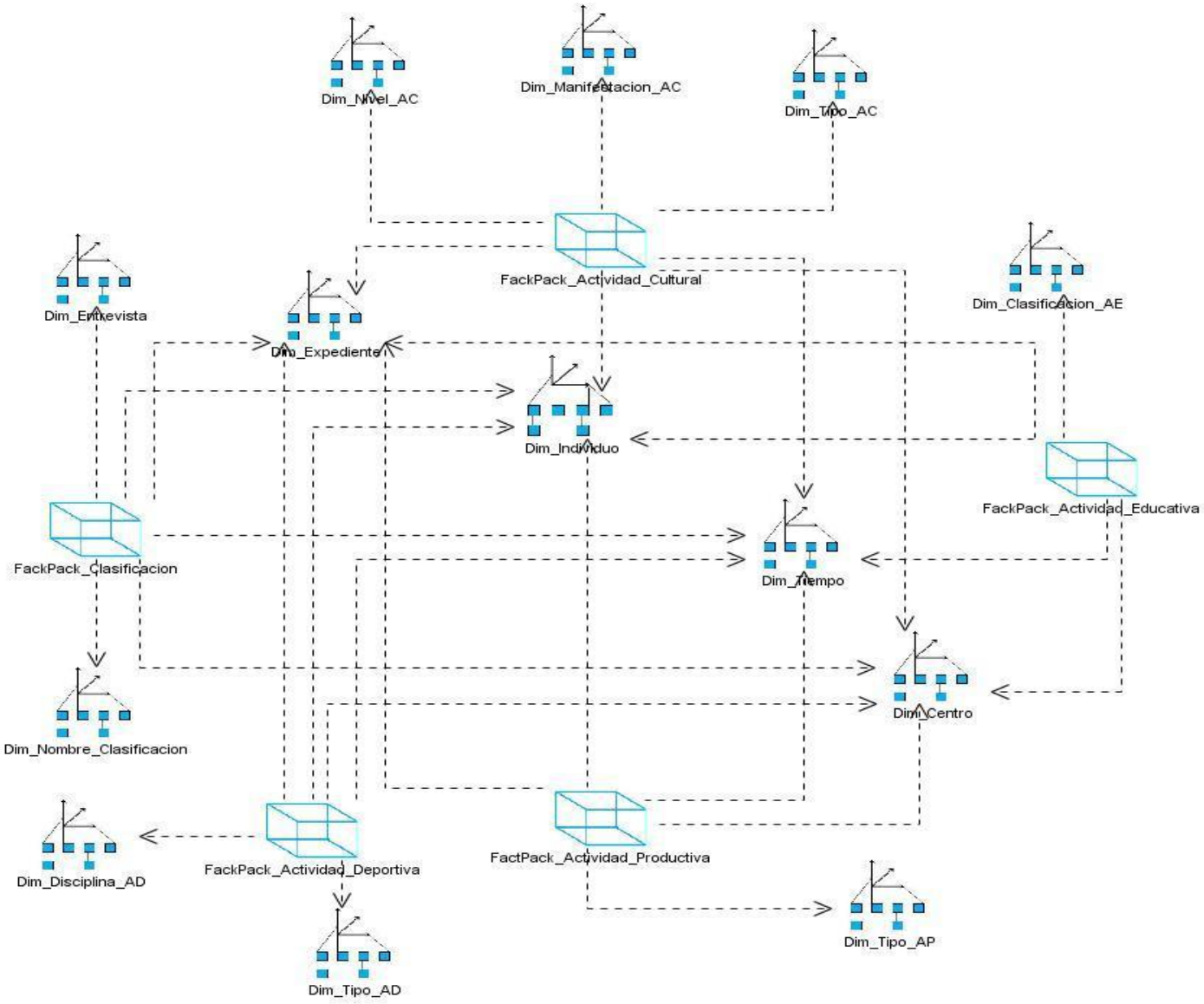
Anexo 10. Esquema conceptual del almacén de datos. Schema_Actividad_Educativa (Nivel 3)



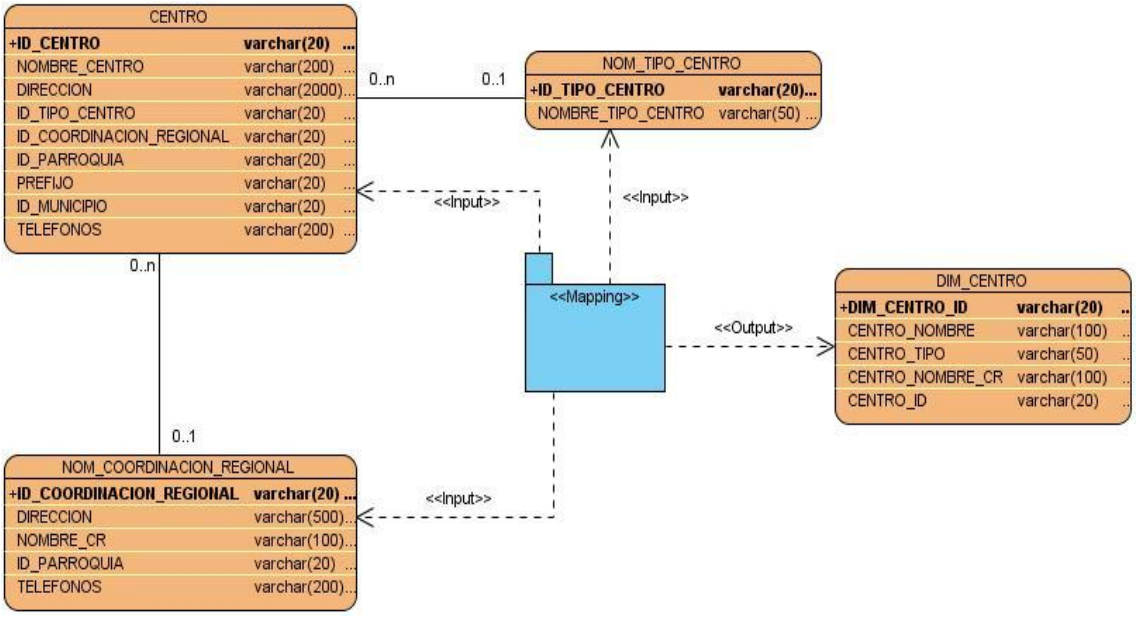
Anexo 11. Esquema conceptual del almacén de datos. Schema_Actividad_Productiva (Nivel 3)



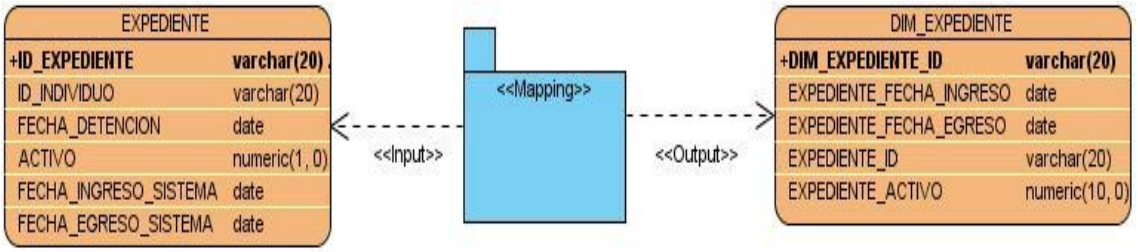
Anexo 12. Vista Global



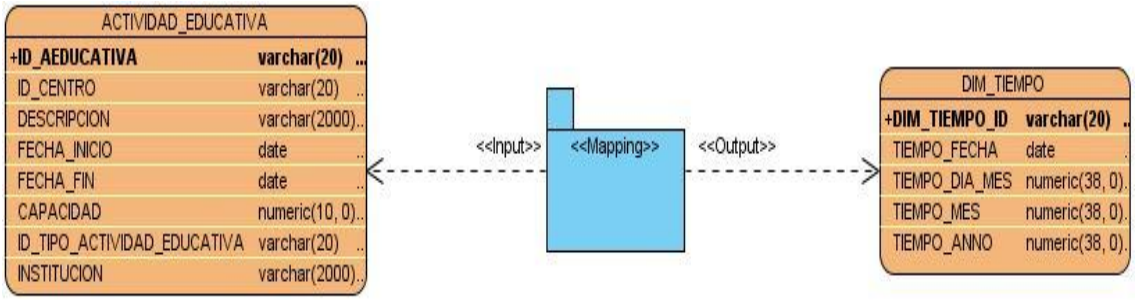
Anexo 13. Mapeo Centro (Nivel 2)



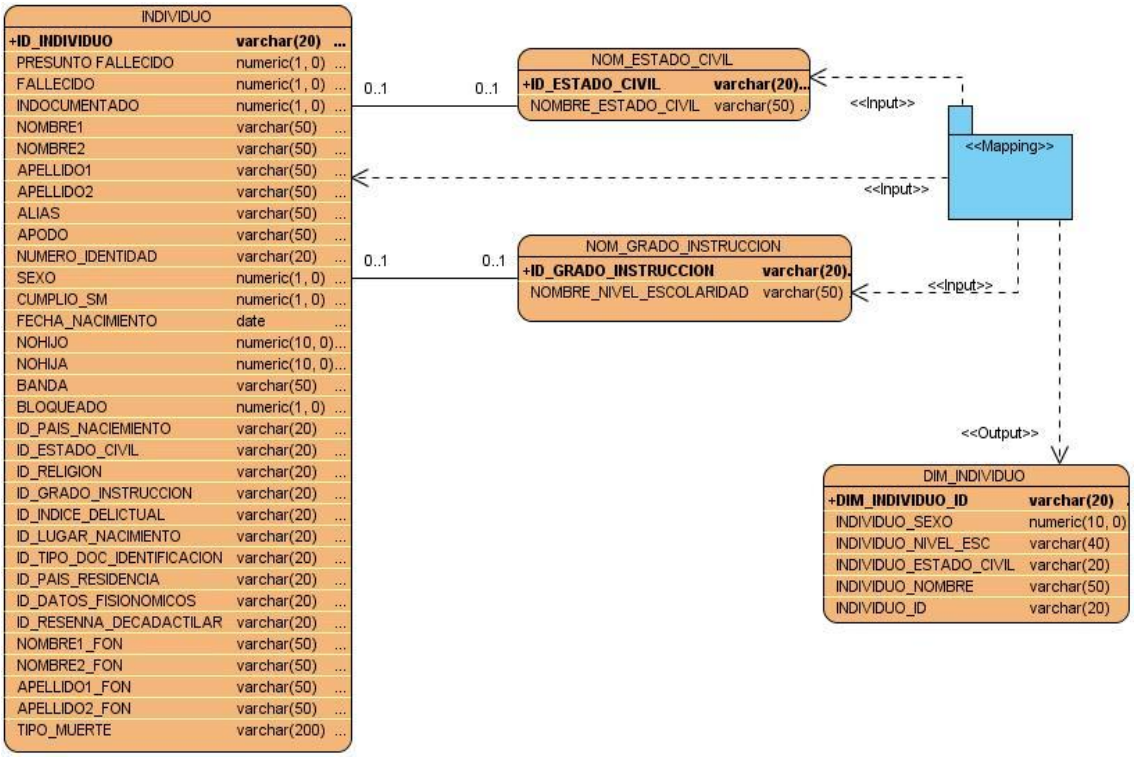
Anexo 14. Mapeo Expediente (Nivel 2)



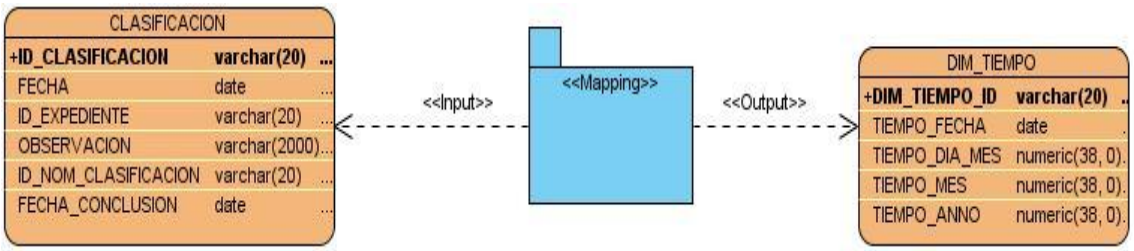
Mapeo 15. Mapeo Tiempo_Actividad_Educativa (Nivel 2)



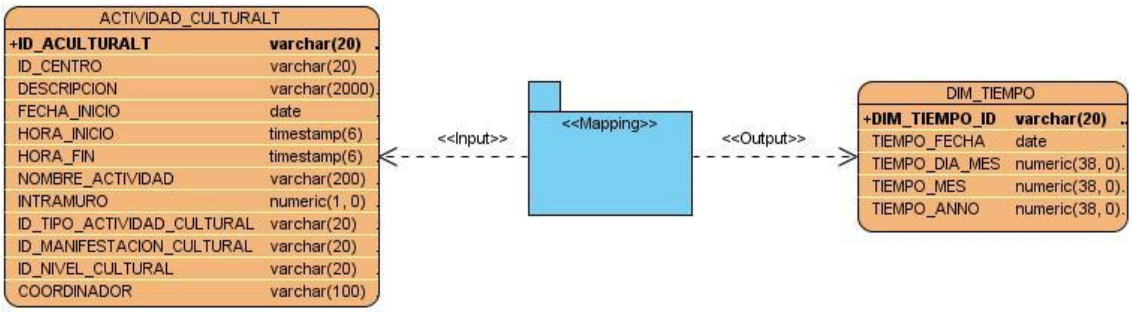
Anexo 16. Mapeo Individuo (Nivel 2)



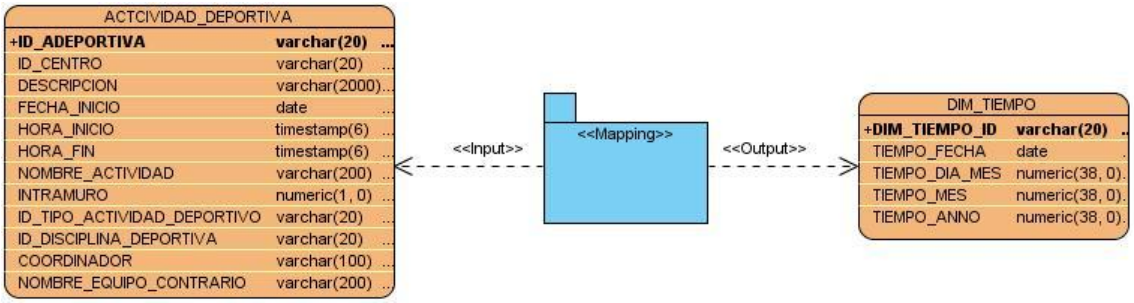
Anexo 17. Mapeo Tiempo_Clasificacion (Nivel 2)



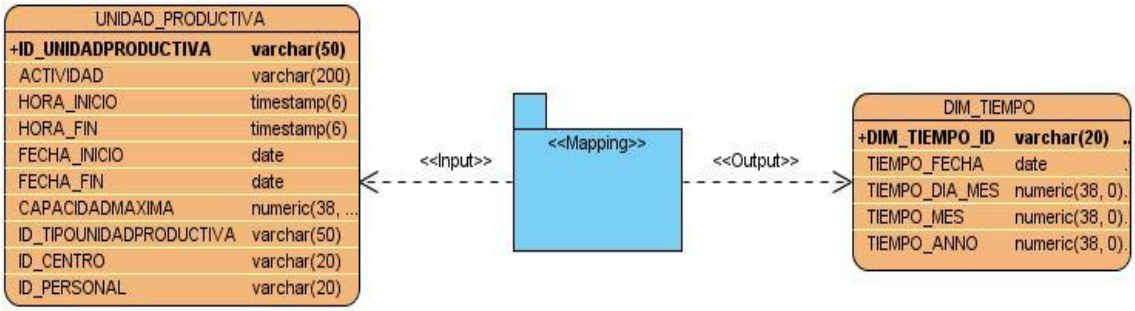
Anexo 18. Mapeo Tiempo_Actividad_Cultural (Nivel 2)



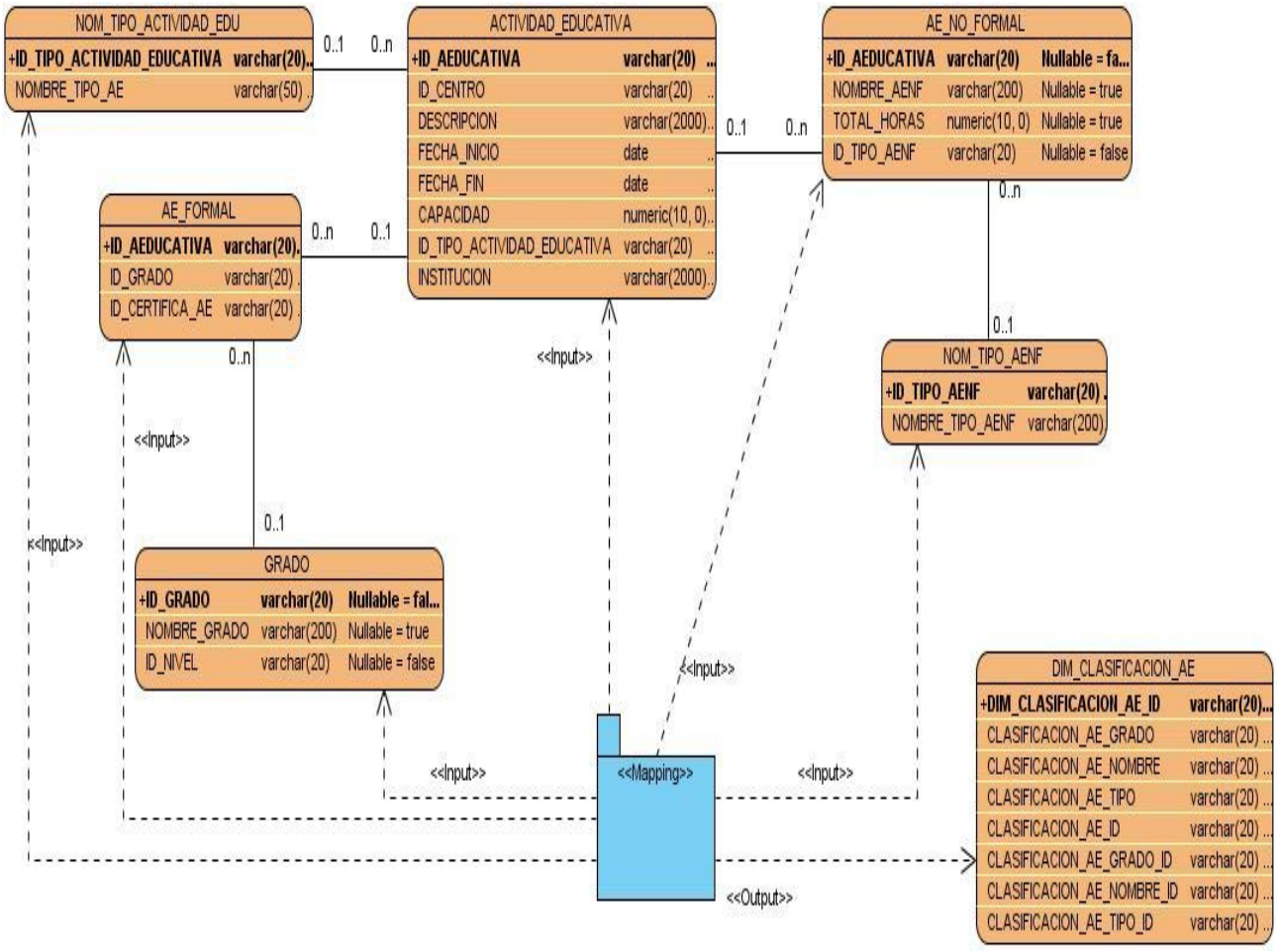
Anexo 19. Mapeo Tiempo_Actividad_Deportiva(Nivel 2)



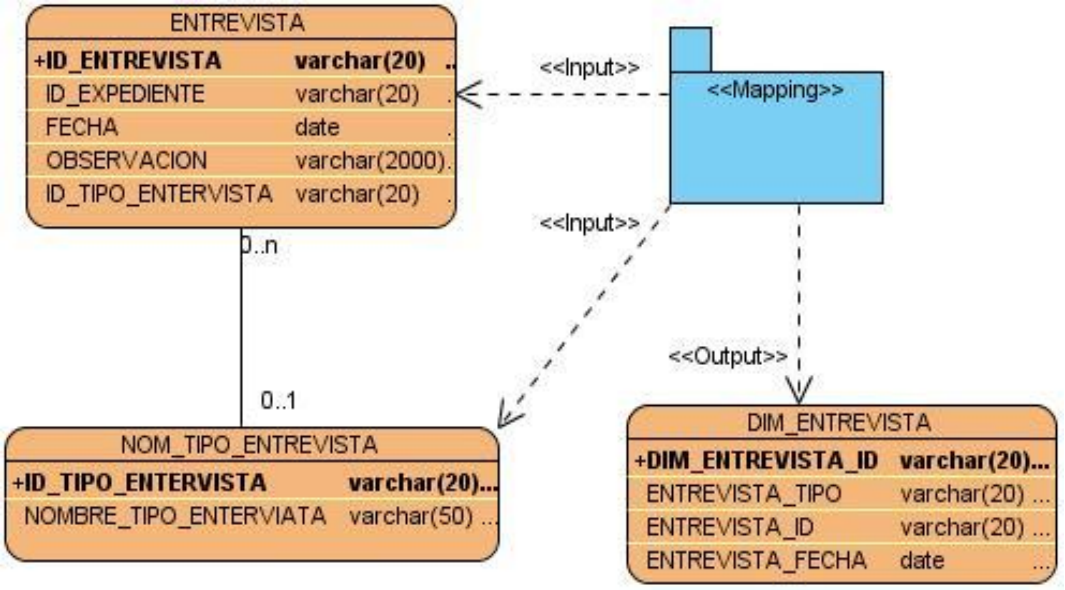
Anexo 20. Mapeo Tiempo_Actividad_Productiva(Nivel 2)



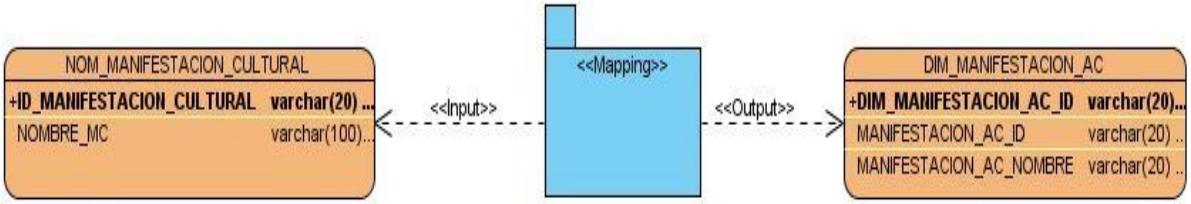
Anexo 21. Mapeo Clasificación_AE(Nivel 2)



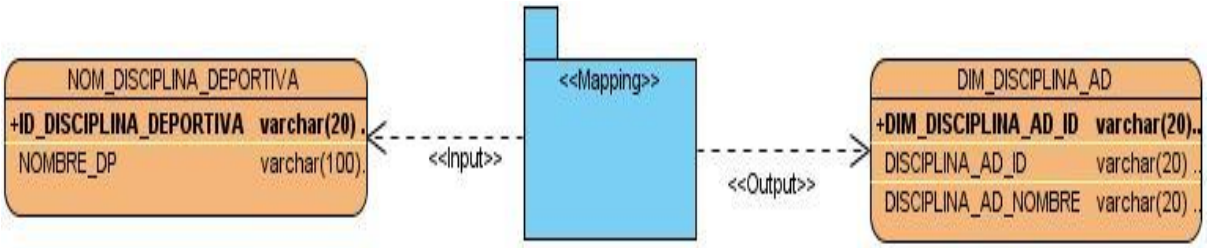
Anexo 22. Mapeo Entrevista (Nivel 2)



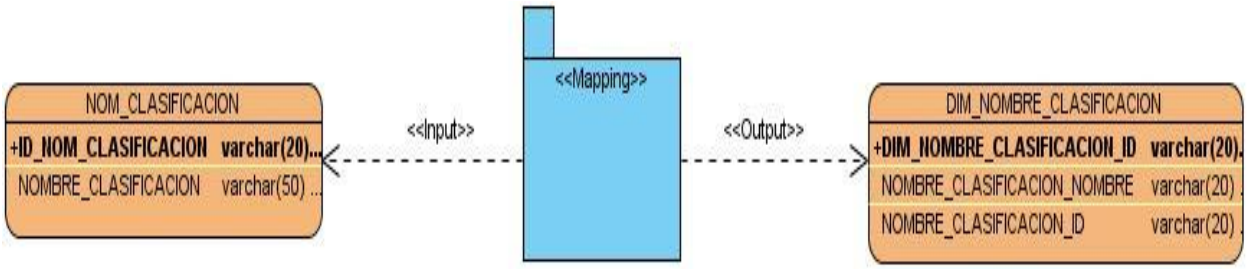
Anexo 23. Mapeo Manifestacion_AC(Nivel 2)



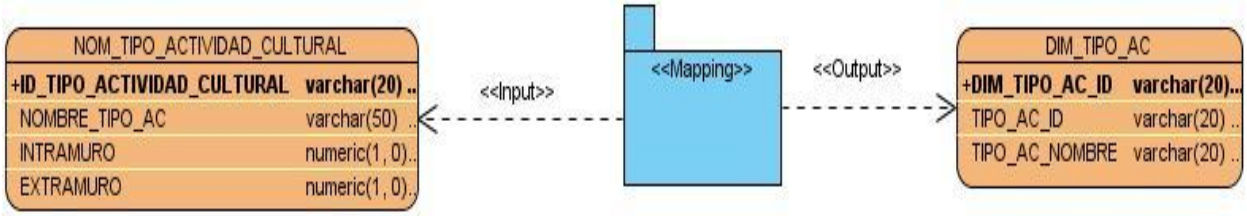
Anexo 24. Mapeo Disciplina_AD (Nivel 2)



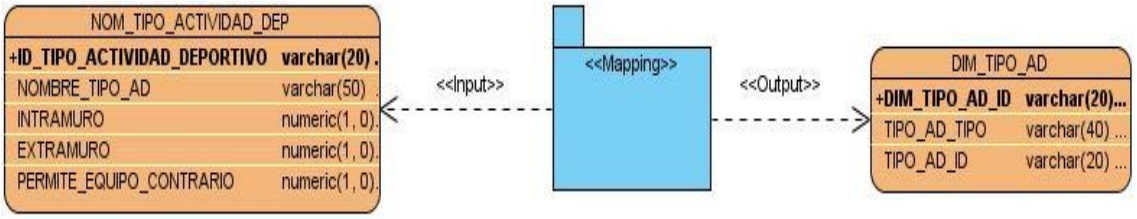
Anexo 25. Mapeo Nombre_Clasificacion (Nivel 2)



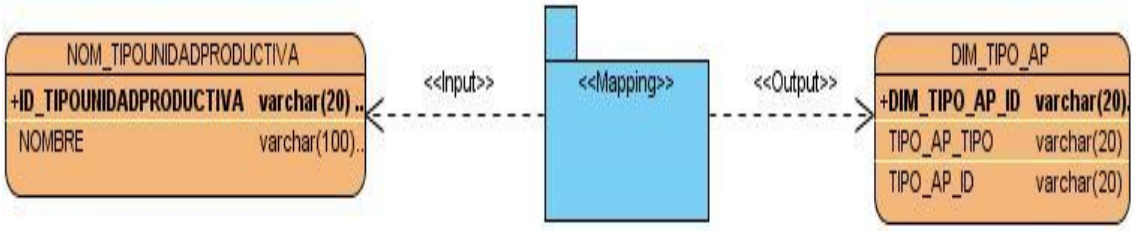
Anexo 26. Mapeo Tipo_AC (Nivel 2)



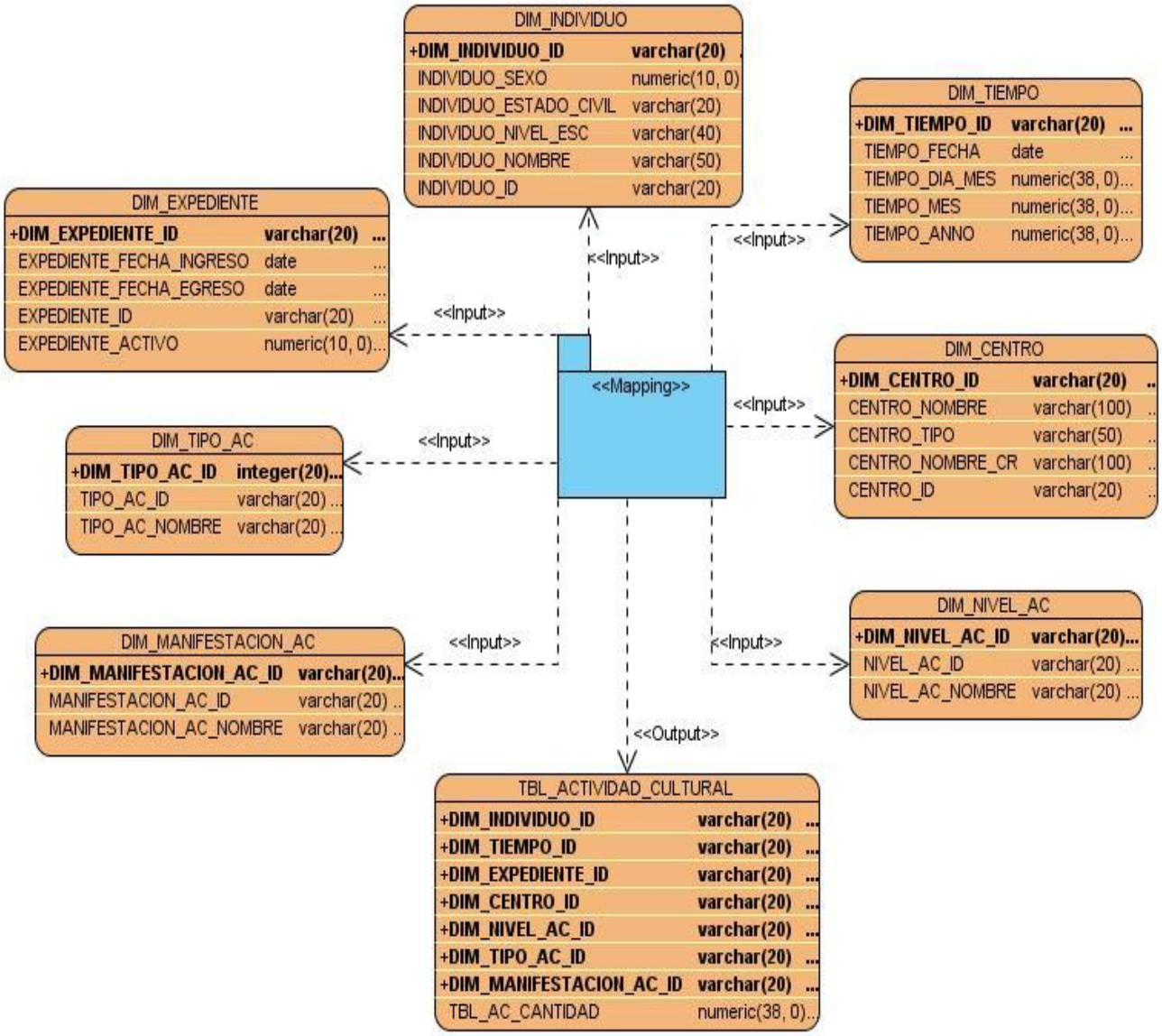
Anexo 27 Mapeo Tipo_AD (Nivel 2)



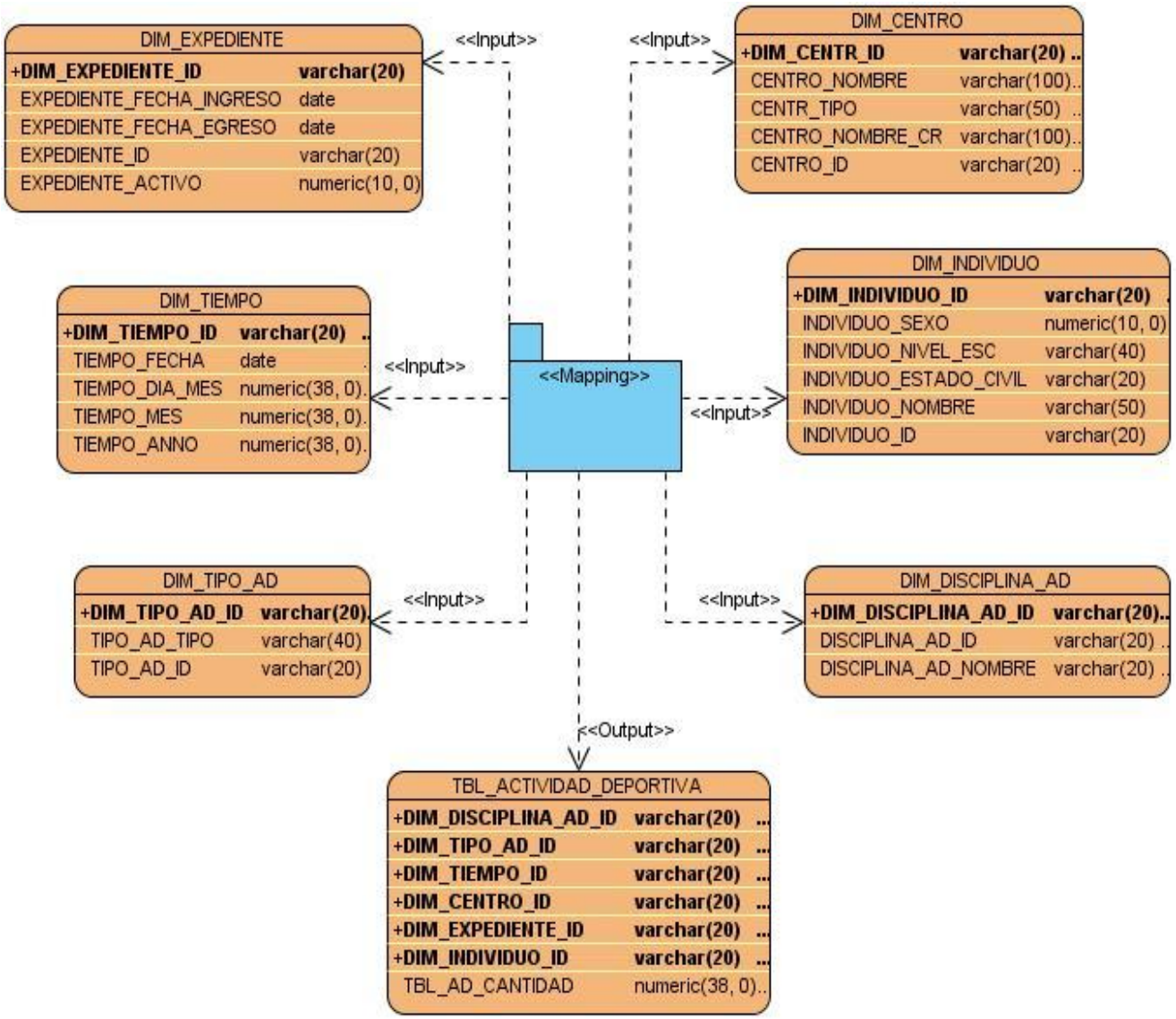
Anexo 28. Mapeo Tipo_AP (Nivel 2)



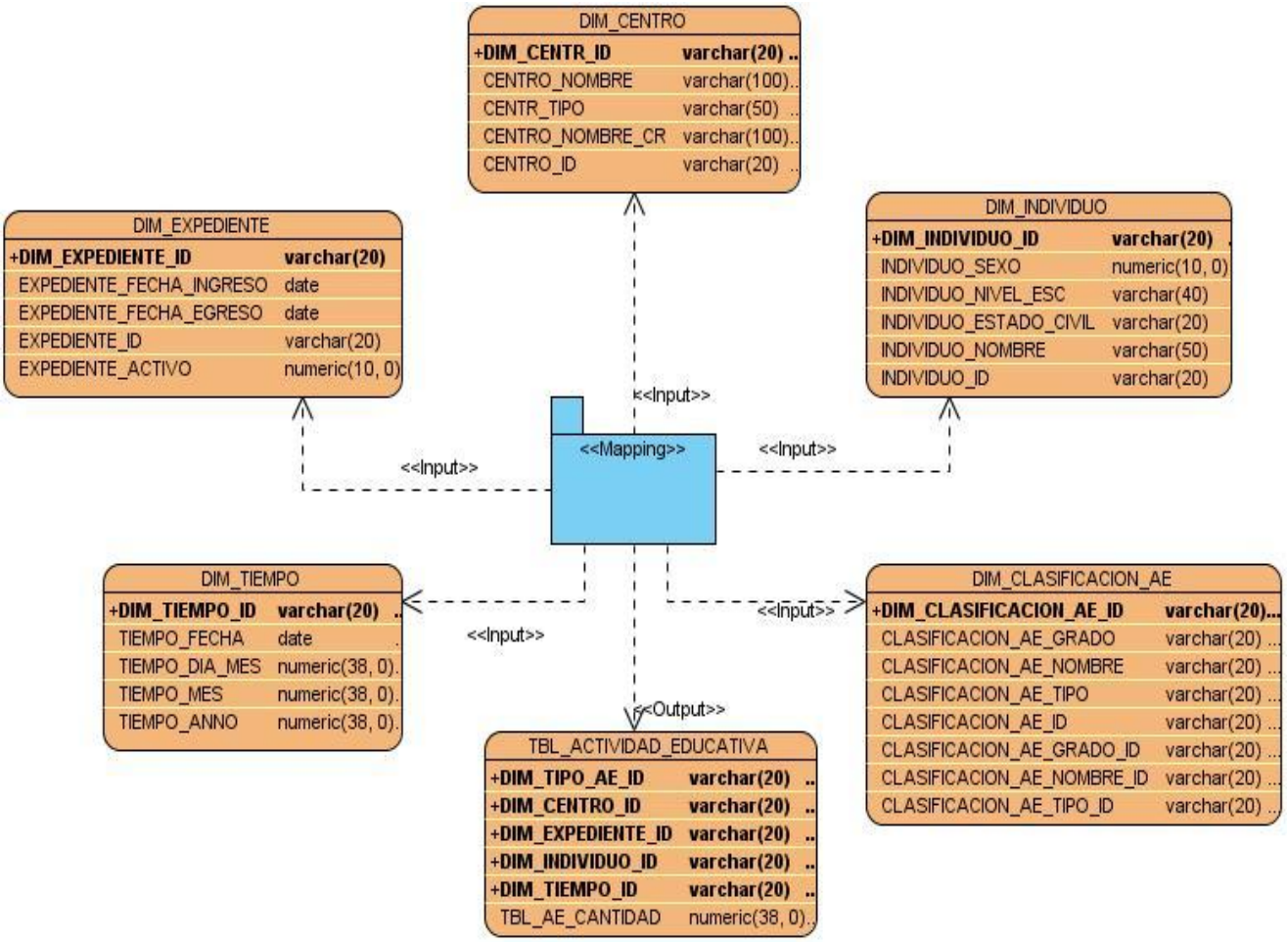
Anexo 29. Mapeo Actividad_Cultural (Nivel 2)



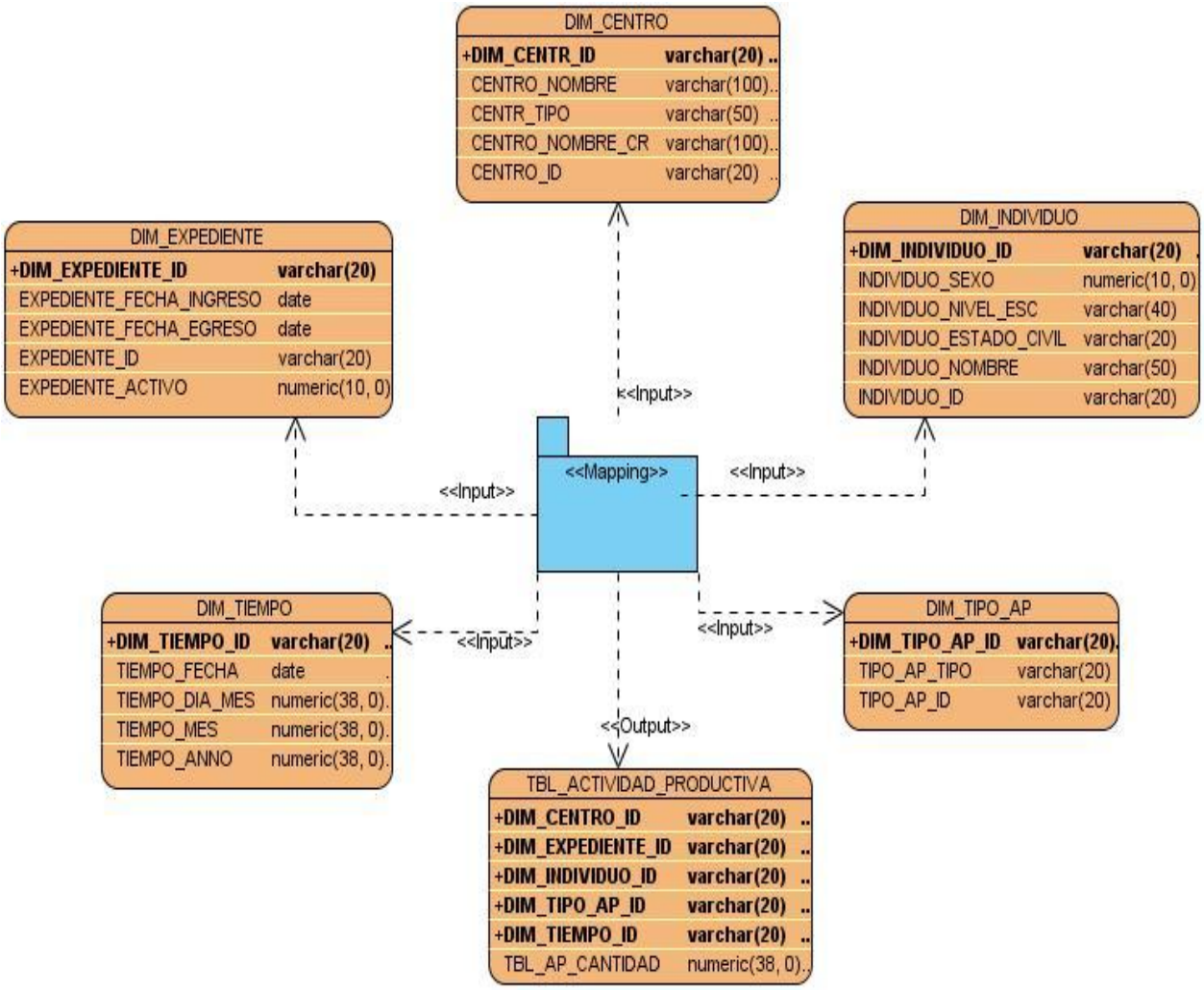
Anexo 30. Mapeo Actividad_Deportiva (Nivel 2)



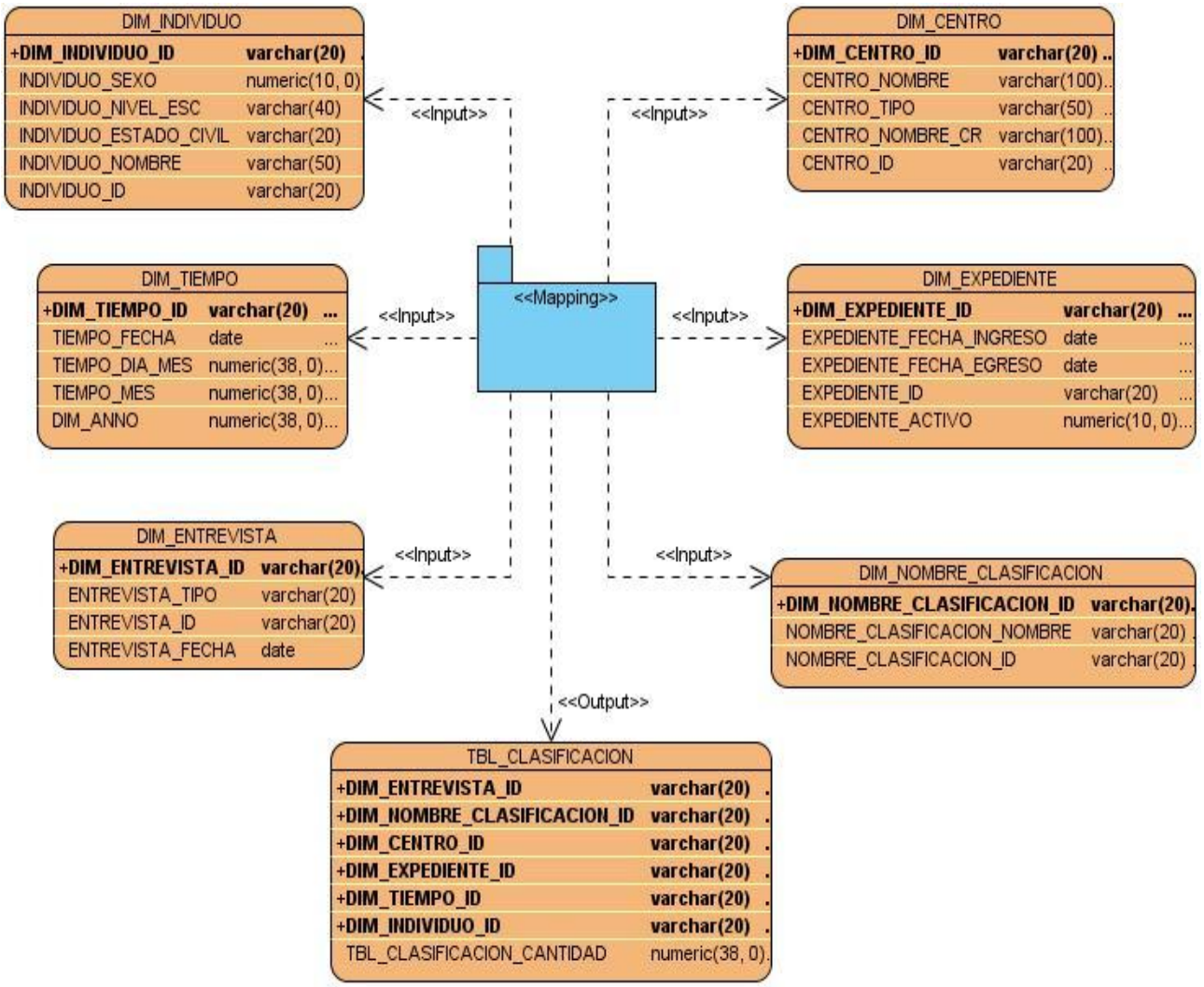
Anexo 31. Mapeo Actividad_Educativa (Nivel 2)



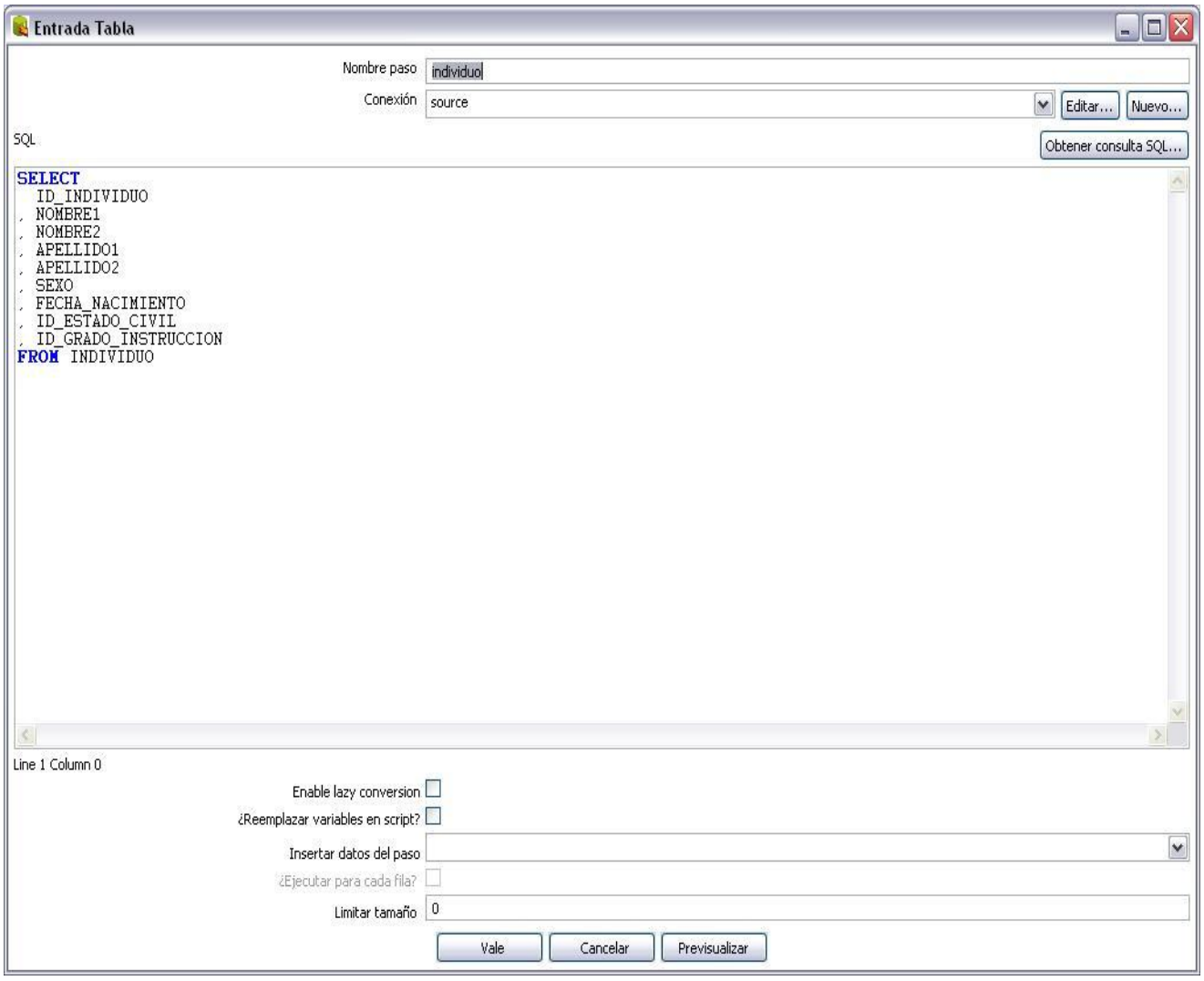
Anexo 32. Mapeo Actividad_Productiva (Nivel 2)



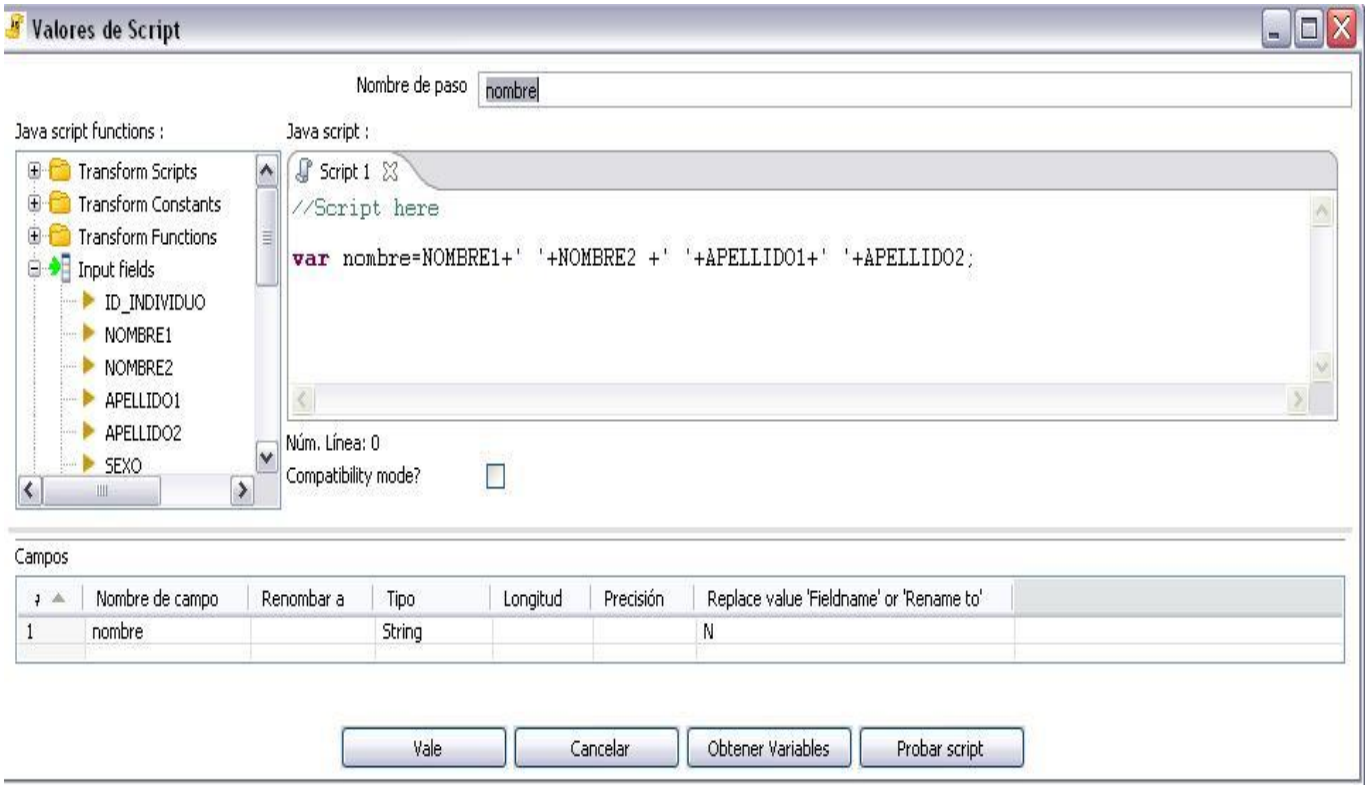
Anexo 33. Mapeo Clasificación (Nivel 2)



Anexo 34. Paso Entrada Tabla Dim_Individuo



Anexo 35. Paso Valor de Java Script Modificado



Anexo 36. Paso Búsqueda en Base de Datos. Estado_Civil. Dim_Individuo