Universidad de las Ciencias Informáticas

Facultad 3

Mercado de Datos para el subsistema Protección de los Derechos Ciudadanos (PDC) del Sistema de Informatización de la Gestión de las Fiscalías (SIGEF).

TRABAJO DE DIPLOMA PARA OPTAR POR EL TÍTULO DE INGENIERO EN CIENCIAS INFORMÁTICAS

Autor: Katherine Zamora Mena

Tutores: Ing. Vlamir Rodríguez Fernández

Co- Tutor: Ing. Manuel Álvarez Alonso

La Habana, 2013.

"Año 55 de la Revolución.**"**

DECLARACIÓN DE AUTORÍA

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

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

_____ _____

Katherine Zamora Mena

Ing. Vlamir Rodríguez Fernández

Si tú tienes una manzana y yo tengo una manzana, e intercambiamos las manzanas, entonces tanto tú como yo seguiremos teniendo una manzana.
Pero si tú tienes una idea y yo tengo una idea, e intercambiamos ideas, entonces ambos tendremos dos ideas.

George Bernard Shaw

Quiero agradecerles a mi mamá y mi papá por haber estado siempre ahí y apoyarme siempre, por ser tan comprensivos, respetar mis decisiones y guiarme por el camino correcto.

A mi hermana que sabe que la adoro, y por darme la alegría de tener unos sobrinos hermosos. A mis primos Dayanis, Dayan, Dairon y Vladimir que para mí son como mis hermanos.

A mis tías Tania y Nancy.

A mi amiga desde el Pre-universitario Yenni que siempre ha estado ahí desde eso entonces.

A mis hermanas desde primer año en la UCI Liliam y Aliuska, a mis amigas Liliana, Ayle y Danay.

A mis compañeras de apartamento Doris, Arletys, Leidys y Arianna.

A mis amigos Polanco, Osmin y Yoandri por ser mis consejeros sentimentales.

A Alexander por los años que compartimos juntos.

A mi tutor Vlamir y mi co-totor Manuel.

A todo aquel que de una forma u otra marco mi paso por esta universidad.

Dedico esta tesis a mi padre Iván Zamora por ser mi guía en todo momento.

A mi mamita Georgina Mena por tanto amor y comprensión.

A mi hermana Doreinis Rodríguez por ser mi apoyo en los momentos difíciles.

A mi abuelita Flor Celeste que no conocí, a mi abuelo José Ramón y a mis abuelos Eulogio y María E. que ya no están entre nosotros.

Y a los niños más importantes de mi vida David, Diana, Dailier y Héctor Luis.

Resumen

Las personas encargadas de la toma de decisiones en las empresas han comenzado a ver la información no solo como un historial de la conducción de la empresa, sino como como un factor que alimenta el negocio y que puede determinar el éxito o el fracaso del mismo. Es por ello que las empresas deben organizar los datos de forma tal que se logre con la consulta de estos ayudar a la toma de decisiones, para la cual una estructura adecuada de almacenamiento es una necesidad esencial.

El presente trabajo permitirá el desarrollo de un mercado de datos para proporcionar el acceso a información actual y datos históricos del subsistema Protección de los Derechos Ciudadanos, permitiendo su análisis desde diversas perspectivas y con grandes velocidades de respuesta.

Este mercado de datos Surge por la necesidad que presenta la Fiscalía General de la República (FGR) de poder almacenar datos de manera eficiente para su explotación y análisis. Su construcción está basada en la metodología de Kimball, la estructura lógica propuesta, el diseño y la implementación son consecuentes con ésta. Se realizan pruebas para validar la capacidad, rendimiento y conectividad del mercado de datos en consonancia los requerimientos definidos por el cliente.

Índice

Índice de figuras

Índice de tablas

Introducción

Los datos son cifras, instrucciones que se tienen, aisladas entre sí, sin seguir una organización o un orden específico, recaudados para un determinado fin. Los datos por si solos no representan información relevante, pero realizado un proceso de ordenamiento dejarían de ser símbolos y convertirse en información que puede ser leída, interpretada y compartida.

En la actualidad las empresas manejan gran cantidad de datos generados por las operaciones que realizan diariamente. Estos datos son almacenados por sistemas de información, que pueden ser transaccionales, o sistemas de soporte a la toma de decisiones. En el caso de los segundos estos mantienen un historial de las transacciones realizadas o decisiones tomadas por las empresas. Sin la existencia de una herramienta que apoye la toma de decisiones, los datos pasan a ser prácticamente inservibles.

Las personas encargadas de la toma de decisiones en las empresas han comenzado a ver la información no solo como un historial de la conducción de la empresa, sino como como un factor que alimenta el negocio para poder determinar el éxito o el fracaso del mismo. Es por ello que las empresas deben organizar los datos de forma tal que se logre con la consulta de estos ayudar a la toma de decisiones.

Existen varias formas de relacionar estos datos, entre ellas se encuentran los sistemas Procesamiento de Transacciones en Línea, OLTP por sus siglas en inglés (On-line Transaction Processing) y los sistemas de Procesamiento de Analítico en Línea, OLAP por sus siglas en inglés (On-Line Analytical Processing). La diferencia más importante entre los sistemas OLTP y los sistemas OLAP es la organización de los datos en los sistemas, o el modelo de datos.

Los almacenes de datos son diseñados para realizar eficientemente la extracción, procesamiento y presentación de la información para el análisis y la toma de decisiones. Estos se diferencian de las base de datos transaccionales en su estructura, funcionamiento, rendimiento y propósito. Por ello es necesario contar con un modelo de datos apropiado, el modelo de datos multidimensional es una buena opción para las tecnologías OLAP.

Las herramientas OLAP han sido creadas en función a bases de datos multidimensionales (o Cubos OLAP), que permiten procesar grandes volúmenes de

información, en campos bien definidos, y con un acceso inmediato a los datos para su consulta y posterior análisis.

Cuba ha integrado conocimientos propios con experiencia en el campo internacional de la informática para enfrentar los problemas que diariamente se presentan en la sociedad. Debido a esto muchas de las empresas han decidido digitalizar la información que poseen, para lograr un mayor resultado en su desempeño.

La Universidad de las Ciencias Informáticas (UCI) tiene la misión de informatizar la sociedad cubana y como resultado en mutuo acuerdo con la Fiscalía General de la República (FGR) se ha creado el proyecto Fiscalía, que es el encargado de informatizar los procesos de las fiscalías. La FGR es el órgano encargado de vigilar por el estricto cumplimiento de la constitución, las leyes y las disposiciones legales, actualmente, a causa del déficit de fiscales en el país, la ejecución y control de sus procesos se demora más de lo previsto, por lo que desarrollar un sistema que mejore en este caso el proceso de ejecución y control de estos procesos llevados a cabo en las fiscalías contribuirá a mejorar el rendimiento de los fiscales y a disminuir los términos.

Entre los subsistemas del SIGEF se encuentra el subsistema Protección a los Derechos Ciudadanos (PDC) que tiene como objetivo principal informatizar los procesos: protección a menores, revisiones laborales, revisiones penales y atención a la población en los que interviene el fiscal, incluyendo los recursos que pueden interponer las partes ante inconformidad, la fiscalía y específicamente la dirección del departamento PDC necesita poseer un repositorio de datos histórico que le permita un acceso rápido a los mismos.

El sistema transaccional existente no facilita realizar la búsqueda de información histórica con agilidad debido a la forma en que está organizada esta, además tampoco permite realizar comparaciones del comportamiento de un dato determinado en el tiempo, siendo esto muy útil para ayudar a la toma de decisiones.

Dada la situación anteriormente descrita, se plantea el siguiente **problema a resolver:** ¿Cómo agilizar el proceso del análisis de los datos generados en el subsistema Protección de los Derechos Ciudadanos del Sistema de Informatización de la Gestión de las Fiscalías para apoyar la toma de decisiones de los directivos de las fiscalías?

El **objeto de estudio** consiste en los Sistemas de Apoyo a la toma de Decisiones.

El **campo de acción** define la gestión de los Mercados de Datos.

Presentando como **idea a defender**: Si se desarrolla un Mercado de datos para el subsistema Protección de los Derechos Ciudadanos del Sistema de Informatización de la Gestión de las Fiscalías se facilitará el análisis de los datos generados en el subsistema y se podrá ayudar con este a la toma de decisiones.

El **objetivo de la investigación es:** desarrollar un Mercado de datos para el subsistema Protección de los Derechos Ciudadanos del Sistema de Informatización de la Gestión de las Fiscalías para favorecer el proceso de toma de decisiones.

Para cumplir el objetivo general del trabajo se trazan los siguientes **objetivos específicos:**

1. Realizar el marco teórico referencial de la investigación.
2. Identificar las necesidades de información.
3. Diseñar el mercado de Datos.
4. Realizar los procesos de extracción, trasformación y carga.
5. Realizar pruebas a la solución.

Para lograr el cumplimiento a los objetivos específicos se plantean las siguientes **tareas de investigación**:

1. Estudio del estado del arte sobre los principales sistemas de soporte a la toma de decisiones y sistemas de software que realizan procesos similares.
2. Análisis de los principales requerimientos de información en el subsistema Protección de los Derechos Ciudadanos.
3. Elección de la granularidad del proceso del negocio.
4. Definición de las dimensiones del mercado de datos.
5. Definición de los hechos mensurables asociados a las dimensiones definidas.
6. Estructuración del modelo dimensional.
7. Transformación del modelo dimensional al diseño físico.
8. Definición de las estrategias de extracción, transformación y carga de los datos.

9. Validación de la solución mediante pruebas de software.

Los **métodos científicos** son los procedimientos que son usados para estudiar la realidad existente con el propósito de descubrir su esencia y sus relaciones, los métodos científicos que se utilizarán para darle solución a los objetivos propuestos son:

Métodos teóricos:

Analítico – sintético: este método se utiliza en el estudio de las diferentes bibliografías para sintetizar el conocimiento aprendido.

Modelación: El modelo es una reproducción simplificada de la realidad que permite descubrir nuevas relaciones y cualidades del objeto de estudio. Este método su evidenciará en el diseño del mercado de datos.

El trabajo está estructurado de la siguiente forma:

Capítulo 1: Fundamentación teórica sobre el desarrollo del mercado de datos. En este capítulo se realiza una breve explicación de las metodologías, técnicas y herramientas que se tuvieron en cuenta para dar solución al problema; así como los principales conceptos y características de las mismas.

Capítulo 2: Diseño del mercado de datos y sus etapas. En este capítulo se muestran aspectos relacionados con el modelado y diseño de la solución propuesta.

Capítulo 3: Implementación. En el capítulo se realiza la implementación de las reglas de transformación de la base de datos relacional al mercado de datos, así como las funciones para mostrar la información.

Capítulo 4: Pruebas de validación. En el capítulo se realizarán un conjunto de las pruebas de calidad con el fin de garantizar la calidad de la solución propuesta.

Capítulo 1: Fundamentación teórica sobre el desarrollo del mercado de datos

Introducción

En este capítulo se abordan definiciones y características fundamentales de los almacenes de datos y de los mercados de datos. Se realiza un estudio sobre la situación actual de su uso además de un estudio bibliográfico de las metodologías existentes en el mundo y en nuestro país para el desarrollo de un mercado de datos. También se describen las distintas herramientas y tecnologías utilizadas para la construcción de los mismos.

1.1 Almacenes de Datos

Varios autores han definido los almacenes de datos, entre ellos se encuentran se encuentran William H. Inmon y Ralph Kimball.

William H. Inmon plantea que:

‘’Un Almacén de Datos es una colección de datos orientados a temas, integrados, no-volátiles y variante en el tiempo, organizados para soportar necesidades empresariales’’. (Inmon, 2002)

Ralph Kimball define un almacén de datos como: "una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis". (Kimball, et al., 2008) También fue Kimball quien determinó que "un almacén de datos no es más que la unión de todos los mercados de datos constituyentes. “ (Kimball, et al., 2008)

Un almacén de datos es una herramienta, con datos integrados, orientados a la consulta y el análisis. Los mismos contienen un registro los datos a través del tiempo, además contiene de estos resúmenes, consolidaciones y análisis de su interrelación.

Un almacén de datos se conforma con datos operacionales y se diseña con el propósito de facilitar la toma de decisiones. La información almacenada en él, solo se habilita para consulta.

1.1.1 Características de los almacenes de datos

Los almacenes de datos se caracterizan por los siguientes aspectos:

Integrados: Integra datos recogidos de diferentes sistemas operacionales de la organización, los cuales deben integrarse en una estructura consistente.

Temáticos: los datos son organizados por temas para facilitar su entendimiento y acceso por parte de los usuarios.

Históricos: se tiene información histórica para poder comparar los datos en diferentes períodos del tiempo e identificar las tendencias de estos. Los datos no son actualizados si no que se tiene almacenado los valores que ha tenido a lo largo del tiempo.

No volátiles: luego que es incorporada la información al almacén de datos debe mantenerse invariable, cargándose una vez en el tiempo y no se permite actualizarla. Los datos solo pueden ser leídos, y no modificados.

1.2 Mercado de Datos

Un Mercado de datos es una base de datos especializada en el almacenamiento de los datos de un área de negocio específica, que tiene como objetivo ayudar a que esta pueda tomar decisiones mejores, por lo que su alcance de contenido es limitado.

Un mercado de datos no es más que un pequeño almacén de datos para un área de negocio específico y centrado en un tema.

1.2.1 Características de un mercado de datos

Los mercados de datos o data mart tienen las siguientes características:

- Usuarios limitados.
- Área específica.
- Propósito específico.
- Función de apoyo.
- Fácilmente entendibles y navegables.

1.3 Arquitectura de un almacén de datos

La arquitectura de un AD viene determinada por su situación central como fuente de información para las herramientas de análisis. Su estructura básica incluye:

- Datos operacionales: fuente de datos para el componente de almacenamiento físico.
- Extracción de datos: selección sistemática de datos operacionales usados para poblar el componente de almacenamiento físico.
- Transformación de datos: Procesos para sumarizar y realizar otros cambios en los datos operacionales y para reunir los objetivos de orientación a temas e integración.

- Carga de datos: inserción sistemática de datos en el componente de almacenamiento físico.
- Almacén: almacenamiento físico de datos.
- Herramientas de acceso: herramientas que proveen acceso a los datos.

A continuación se muestra una imagen que ilustra los componentes de la arquitectura.

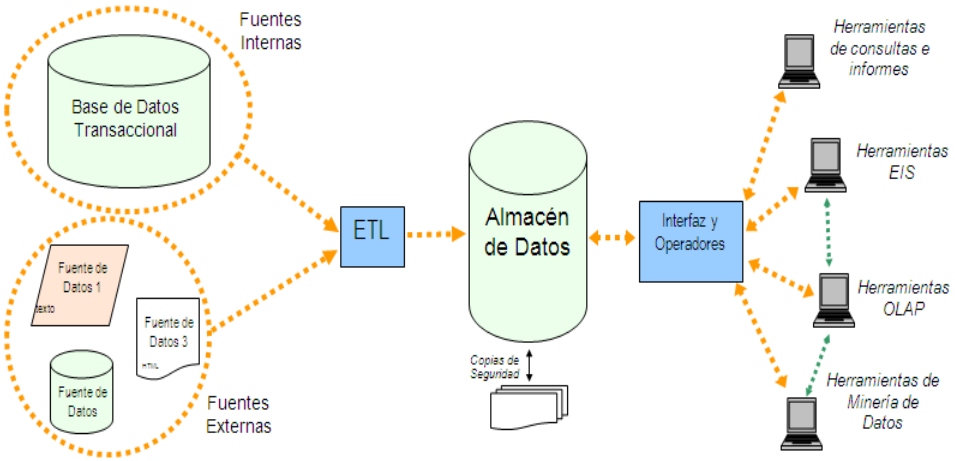


Ilustración 1 Arquitectura de un AD.

1.4 Metodologías de desarrollo para un almacén de datos

Entre las metodologías para la construcción de un almacén de datos se destacan la metodología descendente, propuesta por Bill Inmon y la metodología ascendente propuesta por Ralph Kimball.

La metodología de Kimball propone una arquitectura ascendente, partiendo de la idea de que un almacén de datos es la unión de todos los mercados de datos de una entidad.

Por su parte Inmon define una metodología descendente a la hora de diseñar un almacén de datos, ya que de esta forma se considerarán mejor todos los datos corporativos. En esta metodología los mercados de datos se crean después de haber terminado el almacén de datos completo de la organización.

Después de realizar un estudio de las dos metodologías se puede concluir que la propuesta por Inmon es más compleja y requiere más tiempo para su implementación. Sin embargo Kimball brinda una metodología de fácil comprensión y rápida de implementar por etapas, debido a estas razones y a que el proyecto SIGEF no ha terminado su implementación se decide utilizar la arquitectura para almacenes de

datos propuesta por Kimball para realizar un mercado de datos que ayude al análisis de los datos generados del subsistema PDC.

1.5 Metodología de Kimball (Rivadera, 2010)

La metodología se basa en lo que Kimball denomina ciclo de vida dimensional del negocio. Este ciclo de vida está basado en cuatro principios básicos:

- Centrarse en el negocio: Hay que concentrarse en la identificación de los requerimientos del negocio y su valor asociado, y usar estos esfuerzos para desarrollar relaciones sólidas con el negocio.
- Construir una infraestructura de información adecuada: Diseñar una base de información única, integrada, fácil de usar, de alto rendimiento donde se reflejará la amplia gama de requerimientos de negocio identificados en la empresa.
- Realizar entregas en incrementos significativos: crear el almacén de datos en incrementos entregables en plazos de 6 a 12 meses. Hay que usar el valor de negocio de cada elemento identificado para determinar el orden de aplicación de los incrementos. En esto la metodología se parece a las metodologías ágiles de construcción de software.
- Ofrecer la solución completa: proporcionar todos los elementos necesarios para entregar valor a los usuarios de negocios. Para comenzar, esto significa tener un almacén de datos sólido, bien diseñado, con calidad probada, y accesible. También se deberá entregar herramientas de consulta, aplicaciones para informes y análisis avanzado, capacitación, soporte, sitio web y documentación.

La construcción de un almacén de datos es sumamente compleja, y Kimball propone una metodología que permite simplificar esa complejidad. Las tareas de esta metodología (ciclo de vida) se muestran en la figura 1.2.

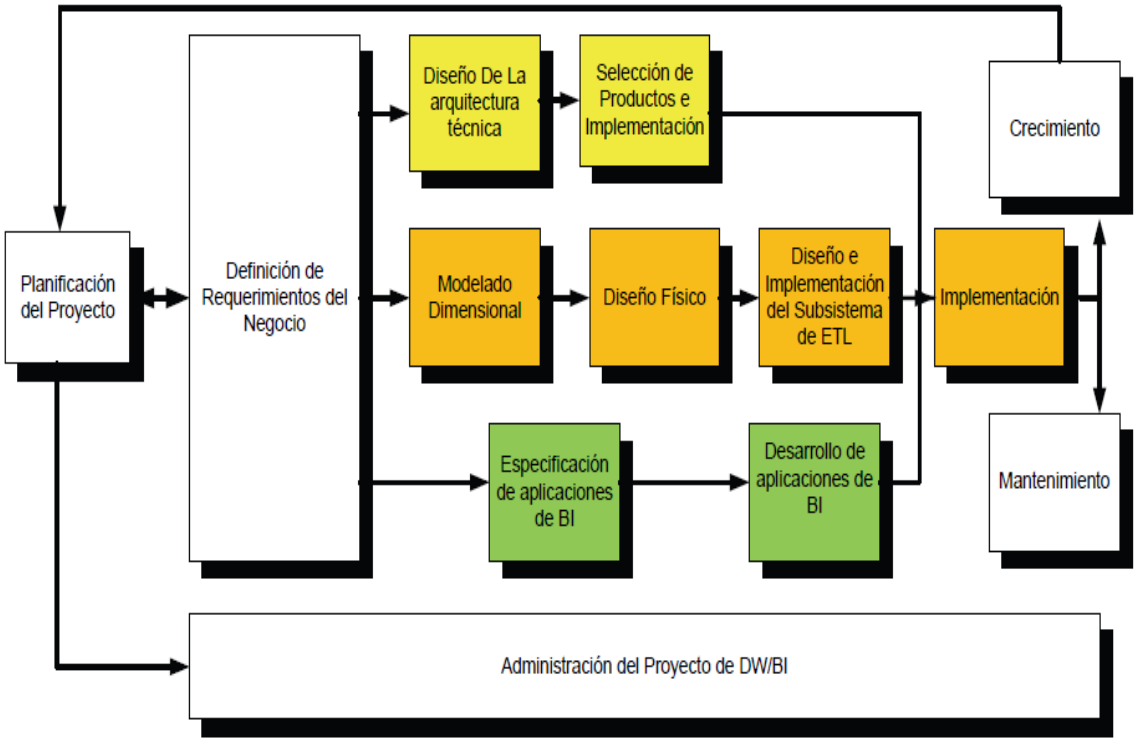


Ilustración 2 Tareas de la metodología de Kimball, denominada Business Dimensional Lifecycle. (Rivadera, 2010)

De la figura 1.2, se puede observar dos cuestiones. Primero, hay que resaltar el rol central de la tarea de definición de requerimientos. Los requerimientos del negocio son el soporte inicial de las tareas subsiguientes. También tiene influencia en el plan de proyecto (nótese la doble fecha entre la caja de definición de requerimientos y la de planificación). En segundo lugar podemos ver tres rutas o caminos que se enfocan en tres diferentes áreas:

- Tecnología (camino superior): Implica tareas relacionadas con software específico.
- Datos (camino del medio): En la misma se diseña e implementa el modelo dimensional, y se desarrolla el subsistema de extracción, transformación y carga.
- Aplicaciones de inteligencia de negocios (camino inferior): En esta ruta se encuentran tareas en las que se diseñan y desarrollan las aplicaciones de negocios para los usuarios finales.

Estas rutas se combinan cuando se instala finalmente el sistema. En la parte inferior de la figura 1.2 se muestra la actividad general de administración del proyecto. A continuación se describe cada una de las tareas.

1. **Planificación**

En este proceso se determina el propósito del almacén de datos, sus objetivos específicos y el alcance del mismo, los principales riesgos y una aproximación inicial a las necesidades de información.

Esta tarea incluye las siguientes acciones típicas de un plan de proyecto:

- Definir el alcance (entender los requerimientos del negocio).
- Identificar las tareas
- Programar las tareas.
- Planificar el uso de los recursos.
- Asignar la carga de trabajo a los recursos.
- Elaboración de un documento final que representa un plan del proyecto.

Además en esta parte se define cómo realizar la administración o gestión de esta subfase que es todo un proyecto en sí mismo, con las siguientes actividades:

- Monitoreo del estado de los procesos y actividades.
- Rastreo de problemas
- Desarrollo de un plan de comunicación comprensiva que direccione la empresa.

2. **Análisis de requerimientos**

La definición de los requerimientos se realiza entrevistando a personas que poseen conocimiento del negocio, para aprender tanto como se pueda sobre este.

Parte del proceso de preparación es definir a quién se debe realmente entrevistar, esto normalmente implica examinar cuidadosamente el organigrama de la organización.

A partir de las entrevistas, se identifican los temas analíticos y procesos de negocio. Los temas analíticos agrupan requerimientos comunes en un tema común.

Por otra parte, a partir del análisis se puede construir un artefacto de la metodología denominado matriz de procesos/dimensiones.

Una dimensión es una forma, vista o criterio por medio de la cual se pueden sumarizar, cruzar o cortar datos numéricos, estos datos que se denominan medidas.

Esta matriz tiene en sus filas los procesos de negocios identificados, y en las columnas, las dimensiones identificadas. Cada X en la intersección de las filas y columnas significa que en el proceso de negocio de la fila seleccionada se identifican las dimensiones propuestas.

3. Modelado Dimensional

La creación de un modelo dimensional es un proceso dinámico y altamente iterativo. El proceso de diseño comienza con un modelo dimensional de alto nivel obtenido a partir de los procesos priorizados de la matriz descrita en el punto anterior.

El proceso iterativo consiste en cuatro pasos:

3.1 Elegir el proceso de negocio

El primer paso es elegir el área a modelar. Esta es una decisión de la dirección que desea desarrollar el mercado de datos, y depende fundamentalmente del análisis de requerimientos y de los temas analíticos anotados en la etapa anterior.

3.2 Establecer el nivel de granularidad

La granularidad significa especificar el nivel de detalle de los datos. La elección de la granularidad depende de los requerimientos del negocio y las características de los datos con los que se cuentan. La sugerencia general es comenzar a diseñar el almacén de datos al mayor nivel de detalle posible, ya que se podría luego realizar agrupamientos al nivel deseado.

3.3 Elegir las dimensiones

Las dimensiones surgen de las discusiones del equipo de desarrollo del mercado de datos. Las tablas de dimensiones tienen un conjunto de atributos (generalmente textuales) que brindan una perspectiva o forma de análisis sobre una medida en una tabla hechos.

3.4 Identificar las tablas de hechos y medidas

El último paso consiste en identificar las medidas que surgen de los procesos de negocios. Una medida es un atributo (campo) de una tabla que se desea analizar, sumarizando o agrupando sus datos, usando los criterios de corte conocidos como dimensiones. Las medidas habitualmente se vinculan con el nivel de granularidad y se

encuentran en tablas que se denominan tablas de hechos. Cada tabla de hechos tiene como atributos una o más medidas de un proceso organizacional, de acuerdo a los requerimientos. Un registro contiene una medida expresada en números, sobre la cual se desea realizar una operación de agregación, en función de una o más dimensiones.

4. **Modelo gráfico de alto nivel**

Para concluir con el proceso dimensional inicial se realiza un gráfico denominado modelo dimensional de alto nivel, donde los óvalos representan las dimensiones y el rectángulo los hechos, las líneas entre ellos representan que el hecho usa la dimensión respectiva. Como ilustra la figura 3.

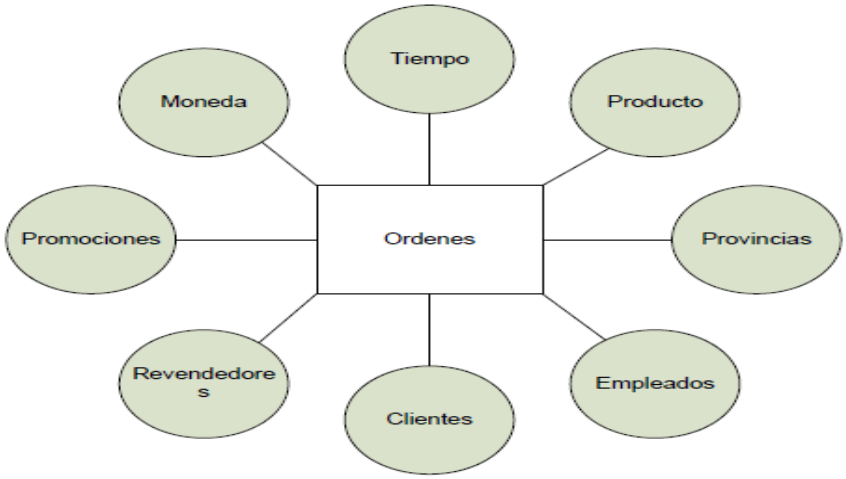


Ilustración 3 Modelo dimensional de alto nivel propuesto por la metodología de Kimball. (Rivadera, 2010)

5. **Identificación de atributos de dimensiones y tablas de hechos:**

La segunda parte de la sesión inicial de diseño consiste en completar cada tabla con una lista de atributos bien formada. Esta lista o grilla se forma colocando en las filas los atributos de la tabla, y en las columnas la siguiente información:

- Características relacionadas con la futura tabla dimensional del almacén de datos, por ejemplo tipo de datos.
- El origen de los datos (por lo general atributos de las tablas transaccionales).
- Reglas de conversión, transformación y carga (ETL), que dicen cómo transformar los datos de las tablas de origen a las del almacén de datos.

6. **Implementar el modelo dimensional detallado**

Este proceso consiste simplemente en completar la información incompleta de los pasos anteriores. El objetivo en general es identificar todos los atributos útiles y sus

ubicaciones, definiciones y reglas de negocios asociadas que especifican cómo se cargan estos datos. Para este cometido se usa la misma planilla del punto anterior.

7. **Prueba del modelo**

Si el modelo ya está estable, lo que se hace habitualmente es probarlo contra los requerimientos del negocio.

8. **Documentos finales**

El producto final es una serie de documentos entre los que se encuentran:

- Modelo de datos inicial de alto nivel.
- Lista de atributos.
- Diagrama de tablas de hechos.
- Definición de campos de medida.
- Diagrama de tablas de dimensiones.
- Descripción de los atributos de las dimensiones.
- Matriz DW.

9. **Diseño físico**

En esta fase, se contestará la siguiente pregunta:

- ¿Cómo convertir el modelo de datos lógico en un modelo de datos físicos en la base de datos relacional?

10. **Diseño del sistema de Extracción, Transformación y Carga (ETL)**

El sistema de Extracción, Transformación y Carga (ETL) es la base sobre la cual se alimenta el almacén de datos. Si el sistema ETL se diseña adecuadamente, puede extraer los datos de los sistemas de origen de datos, aplicar diferentes reglas para aumentar la calidad y consistencia de los mismos, consolidar la información proveniente de distintos sistemas, y finalmente cargar la información en el almacén de datos en un formato acorde para la utilización por parte de las herramientas de análisis.

1.6 Sistemas OLAP

En el mundo de las soluciones para inteligencia de negocios, una de las herramientas más utilizadas por las empresas son las aplicaciones OLAP, ya que las misma han sido creadas en función a bases de datos multidimensionales, que permiten procesar grandes volúmenes de información, en campos bien definidos, y con un acceso inmediato a los datos para su consulta y posterior análisis. (Informática_hoy, 2007-2012)

Las herramientas OLAP proporcionan a las compañías un sistema confiable para procesar datos que luego serán utilizados para llevar a cabo análisis e informes que permitan mejorar las operaciones productivas, tomar decisiones inteligentes y optimizar la competitividad en el mercado. (Informática_hoy, 2007-2012)

Para funcionar, las aplicaciones OLAP utilizan un tipo de base de datos que posee la peculiaridad de ser multidimensional, denominada comúnmente Cubo OLAP. (Informática_hoy, 2007-2012)

1.6.1 Cubos OLAP

Básicamente, el Cubo OLAP, que acuña su nombre por su característica multidimensional, es una estructura conformada por un conjunto de dimensiones, que proporciona respuestas rápidas a consultas analíticas complejas e iterativas utilizadas generalmente para sistemas de ayuda a la toma de decisiones.

Gracias a la incorporación de las bases de datos de tipo multidimensional, y el nacimiento del nuevo concepto Cubo OLAP, las herramientas de soluciones para sistemas inteligencia de negocios han avanzado notablemente en cuanto a las prestaciones que estas aplicaciones brindan a las empresas, donde la información confiable, precisa y en el momento oportuno, son uno de los bienes más preciados.

Un cubo dimensional puede estar formado por los siguientes objetos:

- Indicadores: son sumarizaciones (suma, conteo, promedio, entre otros), efectuadas sobre algún hecho.
- Atributos: son criterios utilizados para analizar los indicadores. Se basan, en los datos de referencia de las tablas de dimensiones. En un cubo, los atributos son los ejes del mismo. Son campos o criterios de análisis, pertenecientes a tablas de dimensiones.

- Jerarquías: Una jerarquía representa una relación lógica entre dos o más atributos, si poseen una relación “padre-hijo”.

1.6.2 Modelos de datos

Hay diversos tipos de implementaciones de la tecnología OLAP, las que varían según el tipo de motor en el que se almacenan los datos. Los cuales se pueden clasificar de la siguiente manera:

ROLAP (Relational On Line Analytic Processing): son sistemas en los cuales los datos se encuentran almacenados en una base de datos relacional. Este tipo de organización física se implementa sobre tecnología relacional, pero disponen de algunas facilidades para mejorar el rendimiento.

MOLAP (OLAP Multidimensional): en estos sistemas se encuentran almacenados los datos en una estructura de datos multidimensional. De manera que la representación externa y la interna coincidan.

HOLAP HOLAP (Hybrid On Line Analytic Processing): el procesamiento analítico híbrido en línea, constituye la unión entre MOLAP y ROLAP, combinando estas dos implementaciones para almacenar algunos datos en un motor relacional y otros en una base de datos multidimensional. (Informática_hoy, 2007-2012)

Se decide que el tipo de almacenamiento será MOLAP porque la estructura de datos almacenada es una estructura multidimensional.

1.7 Integración de los datos.

La integración de los datos consiste en unir los datos que se encuentran en diferentes fuentes en una misma fuente. Esta integración se logra con el proceso ETL.

ETL es el proceso que organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un almacén de datos.

La idea es que una aplicación ETL lea los datos primarios de unas bases de datos de sistemas principales, realice transformación, validación, el proceso cualitativo, filtración y al final escriba los datos en el almacén dejándolos en ese momento disponibles para analizar por los usuarios.

La **extracción** se basa en la acción de obtener la información deseada a partir de los datos almacenados en fuentes externas, la **transformación** es la operación realizada

sobre los datos para que puedan ser cargados en el almacén de datos y la **carga** consiste en almacenar los datos en la base de datos final.

1.8 Base de datos multidimensional

Las bases de datos multidimensional utilizan dos tipos de tablas en sus bases de datos estas son las tablas de hechos y las tablas dimensiones. Además tienen tres variantes posibles de modelamiento: esquema en estrella, esquema constelación y el esquema copo de nieve.

1.8.1 Tabla de hechos

La tabla de hechos es donde las mediciones numéricas del negocio son almacenadas. Cada una de las mediciones es tomada como la intersección de todas las dimensiones. (Kimball, et al., 2008)

1.8.2 Las tablas dimensionales

Las tablas dimensionales son aquellas donde las descripciones textuales de las dimensiones del negocio son almacenadas. Cada una de las descripciones textuales ayuda a describir un miembro de la dimensión respectiva. (Kimball, et al., 2008)

1.8.3 Variantes de modelamiento

Esquema en estrella: posee una gran tabla central (tabla de hechos) y un conjunto de pequeñas tablas acompañantes (tablas dimensiones) presentadas en un modelo radial alrededor de la tabla central.

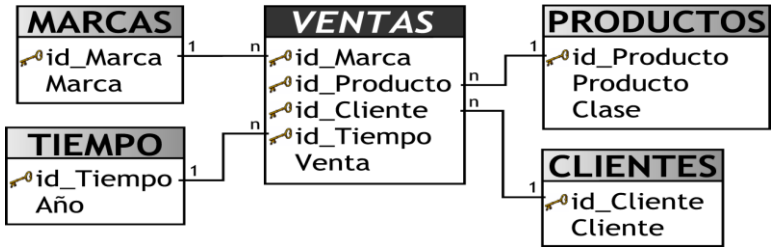


Ilustración 4 Esquema estrella. (Dario, 2009)

Esquema constelación: este modelo está formado por una tabla de hechos principal y por una o más tablas de hechos auxiliares las cuales pueden ser sumarizaciones de la principal. Dichas tablas yacen en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones. (Mavilio, 2009)

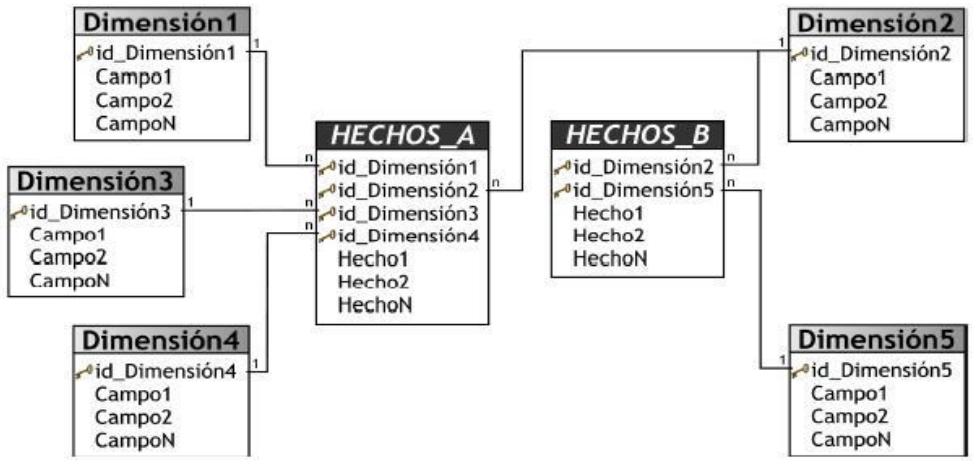


Ilustración 5 Esquema constelación. (Dario, 2009)

Esquema copo de nieve: este esquema representa una extensión del modelo en estrella cuando las dimensiones se organizan en jerarquías de dimensiones. (Mavilio, 2009)

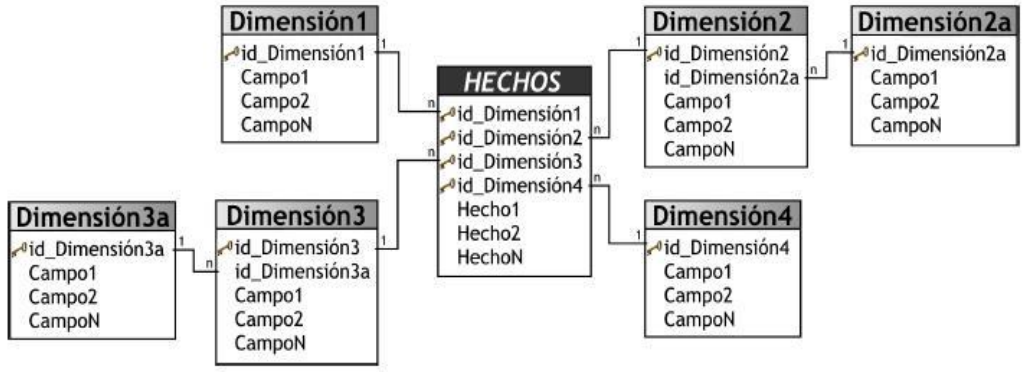


Ilustración 6 Esquema copo de nieve. (Dario, 2009)

1.9 Tendencias

Los almacenes de datos ayudan a las empresas a organizar su información y a tenerla disponible para el análisis, permitiendo ver el comportamiento de esta en diferentes periodos del tiempo por esa razón estas han comenzado a implementarlos para contribuir a su mejor desempeño.

1.9.1 Tendencias actuales de los almacenes de datos en el mundo

El concepto de almacenes de datos no es nuevo existen muchas empresas y organizaciones que han creados almacenes de datos para mejorar la gestión de su información y ayudar a la toma de decisiones entre ellas se encuentran:

La Oficina Nacional de Estadística (ONE) Dominicana tiene un almacén de datos para el Sistema Estadístico Nacional. El almacén de datos es una herramienta electrónica que tiene como objetivo principal disponer de estadísticas precisas y actualizadas para mejorar los servicios al alcance de los tomadores de decisiones, los investigadores y la sociedad en general. El almacén de datos apoya la recolección, procesamiento, análisis y difusión de informaciones estadísticas y ofrece servicios de información basados en modelos analíticos y de minería de datos.

Actualmente, las informaciones disponibles para los usuarios a través del almacén de datos son el Censo Nacional de Población y Vivienda 2002, el Directorio de Establecimientos Económicos 2009, los nacimientos, defunciones, matrimonios y divorcios registrados entre el 2001-2008, las encuestas de Ingresos y Gastos de los Hogares 2007 y las de Hogares de Propósitos Múltiples 2005, 2006 y 2007, entre otras. (ONE , 2011)

DATATUR: creado en el Instituto de Estudios Turísticos (DATATUR) para el análisis y difusión de la información estadística del turismo en España.

La difusión de la información estadística siempre se ha hecho a través de los métodos tradicionales como tablas de datos en formato impreso o Web. La finalidad del proyecto DATATUR, ha sido la de poner al alcance de los analistas de datos estadísticos del turismo, una herramienta que permite analizar esta información y trabajar con ella de un modo más sencillo que con las tradicionales aplicaciones estadísticas y poder además publicar esta información a través de Internet para ponerla al alcance de todo analista que la necesite. (Navarro, et al., 2002)

Almacén de datos para la presentación del servicio público de información estadística del Instituto nacional de Estadísticas, geografía e informática (INEGI) de México: contiene información estadística obtenida de los programas de censos nacionales, encuestas y registros administrativos para la toma de decisiones y la planeación.

Con el objetivo de mejorar los procesos de generación y explotación de información estadística de modo que esta esté disponible en línea de una manera ágil entendible para quien lo requiera con el objetivo de apoyar al servicio público de información estadística en beneficio de los diferentes sectores de la sociedad mexicana.

Fue creado para la mejora del repositorio de información estadística del país para consulta y análisis que permita a los usuarios tomar mejores decisiones como parte

del Sistema Nacional de Información Estadística y Geográfica. (Revista de Arquitectura e Ingeniería, 2010)

1.9.2 Tendencias actuales de los almacenes de datos en Cuba

Cuba no se ha quedado atrás en el desarrollo tecnológico y algunas empresas cubanas se han dedicado a la confección de almacenes de datos que ayudan a mantener un historial de su información y apoyan a la toma de decisiones de las mismas, entre estos se encuentran:

Almacén de Datos para la Gestión Contable de la EMPAI: tiene como objetivo gestionar y organizar homogéneamente la información relevante actual e histórica sobre los indicadores de eficiencia de la gestión contable de la Empresa de Proyectos de Arquitectura e Ingeniería (EMPAI) de Matanzas.

El almacén de datos garantiza mayor disponibilidad y calidad de la información y realiza un análisis dinámico de la misma, facilitando así la toma de decisiones estratégicas de sus directivos, el ahorro de tiempo de los especialistas del área de Contabilidad y de materiales de oficina. (Revista de Arquitectura e Ingeniería, 2010)

La UCI cuenta con el Centro de Tecnología de Gestión de Datos DATEC (Data Tecnology Center), en el que se han desarrollado algunos almacenes de datos como es el de la Oficina Nacional de Estadística (ONE) y la corporación CIMEX. (Martinez, et al., 2010)

La corporación CIMEX se dedica fundamentalmente a la exportación e importación de mercancías. El AD centra su atención en la actividad del comercio, principalmente en la gestión de inventario, permitiendo una gestión de compra–venta eficiente, con una finalidad fundamental: disminuir los costos, sin afectar al cliente, permitiendo prestaciones eficientes y con la calidad requerida, aumentando las ganancias o utilidades de las empresas. (Martinez, et al., 2010)

1.10 Herramientas

1.10.1 Herramienta de inteligencia de negocio

Las herramientas de Inteligencia de Negocios han sido creadas para apoyar la toma decisiones de las empresas o instituciones del estado. Estas muestran una visión general de los procesos que se llevan a cabo en las empresas, facilitan el análisis y la presentación de los datos.

Entre estas herramientas de inteligencia de negocio se pueden encontrar algunas como:

- **Crystal Reports** (2xMil Soluciones, 2012)

Crystal Reports es una herramienta potente a la vez que fácil de usar para el diseño y generación de informes a partir de datos almacenados en una base de datos u otra fuente de información. Es una herramienta de software propietario.

La arquitectura de Crystal Reports.NET gira alrededor del soporte para un tipo de ficheros de formato propietario, que se distingue por la extensión **.RPT** (**report**) y en el que se almacena la definición de los informes. El producto se puede ver como la combinación de tres componentes principales, que son:

Motor de impresión:

El motor de impresión (Crystal Reports Print Engine, CRPE). A pesar de lo que su nombre sugiere, este componente, escrito en código no administrado, no sólo se encarga de lo relacionado con la impresión en papel de los informes, sino además con todo lo que tiene que ver con la ejecución de los mismos, empezando por el acceso a la base de datos para leer la información y continuando con la generación de la imagen de las diferentes páginas para luego volcarlas en pantalla, papel o exportarlas a otros formatos como Adobe PDF o Microsoft Word.

Librerías de código manejado:

Las librerías de código manejado encapsulan la funcionalidad del motor de impresión a través de un conjunto de clases fácilmente accesibles desde aplicaciones escritas en Visual Basic, C# o cualquier otro lenguaje .NET. Estas son usadas para cargar, ejecutar e imprimir los informes.

Diseñador de informes:

El diseñador de informes es el software que presenta la interfaz de usuario a través de la cual un usuario, programador o no, puede crear (“diseñar”) un informe y guardarlo en un fichero .RPT para su posterior reutilización.

- **Oracle Business Intelligence**

Oracle Business Intelligence Suite:

Oracle Business Intelligence Suite es la plataforma más completa para la inteligencia de negocios (BI) disponible en la actualidad, cubriendo un amplio espectro de necesidades de inteligencia de negocios, incluidos los tableros interactivos, alertas e inteligencia proactivas, publicación e informes avanzados, análisis predictivo en tiempo real y análisis de tecnología móvil. Es una herramienta de software propietario. Esta suite está compuesta por varias herramientas entre las que se encuentran:

Oracle BI Publisher, también denominado Oracle XML Publisher, ofrece la solución más eficiente y escalable para informes y publicaciones, disponible para entornos complejos y distribuidos. Oracle BI Publisher brinda una arquitectura central para generar y proporcionar información a los empleados, clientes y socios comerciales, tanto de manera segura como en el formato adecuado.

El software de **Oracle Real-Time Decision** combina los requerimientos comerciales y de información del cliente para hacer la mejor recomendación en cada interacción con el cliente y en cada decisión operacional al adaptarse de manera inteligente la información en constante cambio. Junto con Oracle Business Intelligence Suite y Oracle Fusion Middleware, las empresas pueden aprovechar los conocimientos de las fuentes de datos históricos y en tiempo real para tomar mejores decisiones prácticamente en cualquier situación. (AsiConcultant)

- **Microsoft Business Intelligence**

Microsoft Business Intelligence es un conjunto completo de aplicaciones de servidor, cliente y programador totalmente integrado con 2007 Microsoft Office system, proporciona la información adecuada, en el momento correcto y en el formato idóneo. Además, muestra información fácil de usar directamente en donde trabajan, colaboran y toman decisiones los usuarios. Es una herramienta de software propietario. (Microsoft Corporation, 2012)

A continuación, se muestran algunas características de Microsoft Business Intelligence.

1. **Proporcionar inteligencia empresarial a toda la organización**

Business Intelligence es la base de las necesidades de inteligencia empresarial de la organización. El planeamiento estratégico es más sencillo cuando se usan herramientas conocidas, la administración de la información es más fácil en un entorno de BI centralizado y totalmente integrado y el desarrollo es más rentable cuando se

usa un entorno de desarrollo que se ajusta a los estándares de la industria. (Microsoft Corporation, 2012)

2. **Proporcionar inteligencia empresarial en el nivel empresarial**

Con la plataforma totalmente integrada y completa asistida por SQL Server 2005, Business Intelligence proporciona funcionalidad de extracción, transformación y carga (ETL), procesamiento analítico en línea (OLAP), minería de datos, análisis predictivos y generación de informes, todo en un solo producto. Business Intelligence, que es totalmente escalable, proporciona aprovechamiento, estabilidad, seguridad mejorada y una reducción del costo total de la propiedad. (Microsoft Corporation, 2012)

- **La plataforma Pentaho Open Source Business Intelligence**

La plataforma Open Source Pentaho Business Intelligence (BI) en su versión 3.6.0 cubre las más amplias necesidades de Análisis de los Datos y de los Informes empresariales. Las soluciones de Pentaho están escritas en Java y tienen un ambiente de implementación también basado en Java. Esto hace a Pentaho una solución muy flexible para cubrir una amplia gama de necesidades empresariales tanto las típicas como las sofisticadas y especificas al negocio. (Portada sobre la plataforma Pentaho Open Source Business Intelligence , 2006-2011)

Esta herramienta posee abundante documentación disponible y tiene una interfaz amigable. Esta suite está compuesta por varias herramientas entre las que se encuentran:

Pentaho Analysis:

Pentaho Analysis en su versión 3.0.4, también conocido como Mondrian, es un procesamiento analítico en línea (OLAP) que permite a los usuarios de negocio analizar grandes cantidades de datos en tiempo real. Los usuarios exploran los datos de negocio internándose en la información y la tabulación cruzada con respuestas a la velocidad del pensamiento en consultas analíticas. (summan, 2006-2008)

Pentaho Data Integration:

Pentaho Data Integration en su versión 4.2.1 también conocido como Kettle, ofrece poderosas capacidades de Extracción, Transformación y Carga (ETL) mediante un enfoque innovador, de metadatos. Con un ambiente intuitivo, gráfico, diseño drag and drop y una arquitectura probada, escalable, basada en estándares, la integración de

datos de Pentaho es la elección más demandada por las empresas. (summan, 2006-2008)

Es una de las más antiguas herramientas ETL de código abierto y sin costos de licencia, cuenta con una gran comunidad de usuarios y su interfaz gráfica permite un aumento de la productividad.

Incluye cuatro herramientas (Gravitar, 2012):

Spoon: para diseñar transformaciones ETL usando el entorno gráfico.

PAN: para ejecutar transformaciones diseñadas con spoon.

CHEF: para crear trabajos.

Kitchen: para ejecutar trabajos.

Mondrian Schema Workbench.

WorkBench en su versión 3.2.0 es una herramienta para el desarrollo del esquema del modelo estrella en XML Work Bench desarrollada en Java. Esta herramienta es una interfaz de diseño que permite crear y probar esquemas de cubos OLAP visualmente. Este programa entrega todas las facilidades para poder realizar el modelo lógico del cubo OLAP al cual se le realizarán las consultas. Los modelos de esquemas XML de metadatos se crean en una estructura específica utilizada por el motor de Mondrian. La estructura de estos modelos se pueden considerar de forma de cubos, que utilizan hechos existentes y tablas de dimensiones que se encuentran en el gestor de base de datos. (Díaz, 2010)

Después de análisis de estas herramientas de inteligencia de negocio se concluye que las herramientas Crystal Reports, Oracle Business Intelligence Suite y Microsoft Business Intelligence tienen como principal desventaja que son de software propietario y por tanto se decide que la herramienta a utilizar seria Pentaho Open Source Business Intelligence por ser multiplataforma y de código abierto, y además por la necesidad de tener una soberanía tecnológica.

1.10.2 Sistema gestor de base de datos

Los Sistemas Gestores de Base de Datos (SGBD) están diseñados para gestionar grandes bloques de información. Un SGBD es una aplicación que permite a los usuarios definir, crear, mantener la base de datos y proporcionar un acceso controlado.

PostgreSQL

PostgreSQL es un sistema de gestión de bases de datos objeto-relacional, distribuido bajo licencia BSD y con su código fuente disponible libremente. Es el sistema de gestión de bases de datos de código abierto más potente del mercado y en sus últimas versiones no tiene nada que envidiarle a otras bases de datos comerciales.

PostgreSQL utiliza un modelo cliente/servidor y usa multiprocesos en vez de multihilos para garantizar la estabilidad del sistema. Un fallo en uno de los procesos no afectará el resto y el sistema continuará funcionando.

PostgreSQL 9.1 tiene características técnicas que lo hacen una de las bases de datos más potentes y robustas del mercado. Las características que más se han tenido en cuenta durante su desarrollo son la estabilidad, potencia, robustez, facilidad de administración e implementación de estándares. PostgreSQL funciona muy bien con grandes cantidades de datos y una alta concurrencia de usuarios accediendo a la vez al sistema. (Martinez, 2009-2012)

1.10.3 Administrador de base de datos

PgAdmin III

PgAdmin III en su versión 1.14 es una aplicación gráfica para gestionar el gestor de bases de datos PostgreSQL, siendo la más completa y popular con licencia Open Source. Está escrita en C++ usando la librería gráfica multiplataforma wxWidgets, lo que permite que se pueda usan en Linux, FreeBSD, Solaris, Mac OS X y Windows. Es capaz de gestionar versiones a partir de la PostgreSQL 7.3 ejecutándose en cualquier plataforma, así como versiones comerciales de PostgreSQL como Pervasive Postgres, EnterpriseDB, Mammoth Replicator y SRA PowerGres. (guia-ubuntu, 2008)

PgAdmin III está diseñado para responder a las necesidades de todos los usuarios, desde escribir consultas SQL simples hasta desarrollar bases de datos complejas. El interfaz gráfico soporta todas las características de PostgreSQL y facilita enormemente la administración. La aplicación también incluye un editor SQL con resaltado de sintaxis, un editor de código de la parte del servidor, un agente para lanzar scripts programados, soporte para el motor de replicación Slony-I y mucho más. La conexión al servidor puede hacerse mediante conexión TCP/IP o Unix Domain Sockets (en plataformas *nix), y puede encriptarse mediante SSL para mayor seguridad. (guia-ubuntu, 2008)

1.10.4 Herramienta de modelado

Visual Paradigm 8.0:

Visual Paradigm es una herramienta que sirve para realizar modelado UML. Esta herramienta tiene unas características gráficas muy cómodas que facilitan la realización de los diagramas de modelado. Es factible a la hora de dibujar diagramas de clases, y generar script para diferentes Sistemas Gestores de Bases de Datos. El diagrama que mejor describe las relaciones entre clases y la noción de atributos claves que relacionan entre sí las tablas unas con otras es el Diagrama de Relación de Entidad (ER diagram).

1.10.5 Contenedor Web

Apache Tomcat 5.5.12 es un contenedor de Servlets con un entorno JSP, proyecto de desarrollo Open Source y es publicado bajo la licencia Apache Software License, es un software de código abierto implementado para las tecnologías Java Servlet y Java Server Pages. Se ejecuta en varios sistemas operativos. Es un servidor configurable de diseño modular, con diversidad que permiten garantizar una elevada seguridad y buenas prestaciones.

10.10.6 Herramientas para las pruebas.

EMS Data Generator 2005 for PostgreSQL 2.21.

Es una potente herramienta para la generación de datos de prueba a tablas de bases de datos en PostgreSQL. Permite definir tablas y campos para la generación de datos, establecer rangos de valores, generar campos char por la máscara, obtener listas de valores de las consultas SQL y muchas otras características para generar datos de prueba de forma sencilla y de manera directa. También proporciona aplicación de consola, lo que le permite generar datos en un solo toque mediante el uso de plantillas de generación. Las principales características que posee son. (PostgreSQL, 2005)

- Fácil de usar interfaz de asistente
- Seis idiomas disponibles: Inglés, francés, alemán, italiano, ruso y español.
- Generación de datos a varias tablas de bases de datos diferentes en un host.
- Todos los tipos de datos PostgreSQL, incluyendo Array, direcciones de red y tipos geométricos.
- Los diferentes tipos de generación para cada campo, incluyendo lista, la generación incremental de datos al azar y mucho más.

- Capacidad para utilizar los resultados de consultas SQL como lista de valores para la generación de datos.
- El control automático sobre la integridad referencial para las tablas vinculadas generación de datos.
- Gran variedad de parámetros de generación para cada tipo de campo.
- Capacidad para establecer los valores NULL para cierto porcentaje de los casos.
- Posibilidad de guardar todos los parámetros de generación, creado en la sesión del asistente actual.
- Utilidad de línea de comandos para generar datos utilizando el archivo de plantilla.

Apache JMeter 2.9.

Es un software de código abierto realizado en Java, puede ser utilizado para probar el rendimiento tanto en recursos estáticos y dinámicos (archivos, Servlets, scripts de Perl, Java Objects, bases de datos y consultas, servidores FTP y mucho más). Se puede utilizar para simular una carga pesada en el servidor, de red o un objeto para probar su resistencia o para analizar el rendimiento general bajo diferentes tipos de carga. Se puede utilizar para hacer un análisis gráfico de rendimiento o para probar su servidor / script / comportamiento del objeto bajo carga pesada concurrentes. Sus principales características son (Jmeter, 2013):

- Se puede cargar y probar diferentes tipos de servidores de rendimiento, (Web-HTTP, HTTPS, JABÓN, Base de datos a través de JDBC, LDAP, JMS, Correo-SMTP (S), POP3 (S) e IMAP (S), Comandos nativos o scripts de Shell)
- Portabilidad completa y pureza Java 100%.
- Permite el muestreo simultáneo de muchas discusiones y toma de muestras simultáneas de diferentes funciones de los grupos de hilos separados.
- Su cuidadoso diseño GUI permite un funcionamiento más rápido y los tiempos más precisos.
- Almacenamiento en caché y el análisis / Reproducción de resultados de las pruebas fuera de línea.
- Samplers conectables permiten capacidades de prueba ilimitadas.
- Varias estadísticas de carga pueden ser elegidos con temporizadores enchufables.
- Análisis de datos y plugins de visualización permiten una gran extensibilidad y personalización.

Conclusiones del Capítulo

En el presente capítulo se hizo referencia a los principales conceptos y características de mercado de datos y almacén de datos, además de diferentes criterios de autores en cuanto a estos. Se estudian algunas metodologías existentes y se decide que la más adecuada a las necesidades del proyecto para el desarrollo de un mercado de datos es la metodología propuesta por Kimball. Se realizó un estudio de las tendencias tanto nacionales como internacionales de los almacenes de datos. Y se abordaron los principales elementos que componen los almacenes de datos. Así como las herramientas y tecnologías a utilizar quedando definidas a utilizar para el desarrollo de un mercado de datos para el subsistema PDC del SIGEF las siguientes:

- Pentaho Open Source Business Intelligence.
- Sistema gestor de base de datos: PostgreSQL.
- Administrador de base de datos: PgAdmin III**.**
- Herramienta de modelado**:** Visual Paradigm**.**
- Lenguaje de Modelado Unificado (UML).
- Contenedor Web Apache Tomcat.

Capítulo 2: Diseño del mercado de datos

2.1 Introducción

El proceso de construcción de un mercado de datos comienza por un análisis de los requerimientos donde se identificarán las necesidades de información de los usuarios, luego se realiza un análisis de las fuentes de datos existentes donde se determina el estado general de las mismas. Se tratarán los temas relacionados con el diseño del mercado de datos, así como la arquitectura definida para el mismo. Se identificarán las tablas de hechos y dimensiones además se realizarán las uniones entre las mismas para conformar el modelado del mercado de datos.

2.2 Análisis de las necesidades de información:

La primera etapa del desarrollo de la solución se comienza identificando las necesidades de información de los usuarios, ya que es de gran importancia conocer cuál es la información que necesitan y como es que la desean. Después de realizar un análisis de los reportes del subsistema PDC se identificaron las siguientes necesidades de información:

Para el módulo atención a la población:

Cantidad de personas atendidas en la fiscalía.

Cantidad de personas atendidas en la fiscalía según color de piel según instancia.

Cantidad de personas atendidas en la fiscalía según sexo según instancia.

Cantidad de personas atendidas en la fiscalía según edad según instancia.

Cantidad de personas atendidas por materia según instancia.

Para el módulo quejas y reclamaciones:

Total de reclamaciones por instancia.

Cantidad de reclamaciones por estado según instancia.

Cantidad de reclamaciones por estado según sexo, según instancia.

Cantidad de reclamaciones por estado según edad, según instancia.

Cantidad de reclamaciones por estado según materia, según instancia.

Para el módulo impugnaciones:

Cantidad de impugnaciones por estado de un involucrado dado en una materia por instancia.

Cantidad de impugnaciones por estado de un involucrado por instancia.

Para el módulo menor:

Visitas:

Cantidad de visitas realizadas a un centro determinado.

Cantidad de infracciones detectadas en un centro determinado.

Cantidad de resoluciones desglosadas por centros.

Rollo:

Cantidad de menores con medidas

Cantidad de menores con medidas por edad.

Cantidad de menores con medidas desglosadas por edad.

Cantidad de menores con medidas desglosadas por color de piel.

Cantidad de menores en escuelas de conducta.

Cantidad de menores en escuelas de conducta por categoría.

Cantidad de menores en escuelas de conducta por categoría desglosados por sexo.

Cantidad de menores en escuelas de conducta por categoría por edad.

Cantidad de menores en escuelas de conducta por categoría por color de piel.

Cantidad de menores por centro por sexo.

Cantidad de menores por centro por color de piel.

Cantidad de menores por centro por edad.

Cantidad de menores en centro por motivo.

Expedientes:

Cantidad de expedientes radicados contra menores.

Cantidad expedientes radicados contra menores por motivo.

Cantidad de menores involucrados en expedientes por motivo desglosados por sexo.

Cantidad de menores involucrados en expedientes por motivo desglosados por edad.

Cantidad de menores involucrados en expedientes por motivo desglosados por color de piel.

Total de expedientes revisados por el fiscal.

Cantidad de expedientes revisados por el fiscal por causa de apertura.

Para el módulo Revisiones Penales:

Total de solicitudes de revisión penales recibidas en la fiscalía en una provincia determinada.

Total de solicitudes de revisión penales recibidas en la fiscalía desglosada por mayores de edad y menores en una provincia determinada.

Cantidad de revisiones por estado desglosada por género en una provincia.

Para el módulo Revisiones Laborales:

Total de demandas de revisión laboral recibidas en la fiscalía por provincia.

2.3 Análisis del estado de las fuentes de datos

Para gestionar los datos correspondientes al sistema SIGEF se cuenta con una base de datos relacional y como gestor de base de datos se utiliza PostgreSQL en la versión 9.1.

Como medida organizativa y técnica tomada para asegurar el funcionamiento del sistema SIGEF, los servidores deben estar disponibles las 24 horas del día, por consiguiente la base de datos se deben encontrar en plena disponibilidad. El sistema SIGEF cuenta con una funcionalidad para la réplica de datos, por lo que los servidores de los diferentes niveles se encuentran actualizados.

El subsistema PDC tiene un total de 216 tablas de las cuales 54 son tablas nomencladoras y 162 son tablas de datos. Se pueden señalar como tablas relevantes para el mercado de datos las tablas datencion, dmenores, dinstitucionmenor, dexpedientemenores, dplanteamiento, dresolucion, drolloqueja, drollorevlab, dsolicitud y dvisita. Además son también necesario tablas del subsistema Base entre estos se encuentran: dfiscalia, dpersona y dproceso.

Se estudió la correspondencia de los campos de la base de datos del SIGEF con los del mercado de datos, el cual genero el artefacto modelo lógico de datos. En el anexo 16 se puede observar este modelo.

1.1 Metodología de Kimball

A continuación se le da cumplimiento a los pasos propuestos por la metodología de Kimball para el desarrollo de un mercado de datos.

1.1.1 Planificación

Alcance: El mercado de datos será realizado para los módulos: Atención a la Población, Revisiones Laborales, Revisiones Penales y Protección a Menores del subsistema PDC del SIGEF.

1.1.2 Análisis de requerimientos

Para la realización de esta fase se realiza una reunión con los analistas del SIGEF para determinar las necesidades de información relevantes para las Fiscalías referentes al subsistema PDC.

De esta reunión se seleccionaron los temas analíticos y sus requerimientos que se muestran en la siguiente tabla.

Tema analítico	**Análisis o requerimiento inferido o pedido.**	**Proceso de negocio de soporte.**	**Comentario**
Atención a la población.	Análisis de las personas atendidas en las fiscalías.	Atención a la población.	Por color de piel, sexo, edad, por materia e instancia.
Quejas y reclamaciones	Análisis de las personas que realizan reclamaciones.	Quejas y reclamaciones	Por color de piel, sexo, edad, por materia e instancia.
Impugnaciones	Análisis de las impugnaciones realizadas.	Impugnaciones	Por materia e instancia.

Menores	Análisis de las visitas realizadas a los centros de atención a menores.	Visitas	Por instancia, por centros.
	Análisis de los menores que son tratados por problemas de conducta o delito.	Rollo	Por instancia, por centros, por edad, sexo y color de piel.
	Análisis de los expedientes que han sido abiertos para menores.	Expediente	Por causa, sexo, color de piel y edad.

Tabla 1. Temas analíticos

A partir de este análisis se logró realizar la Matriz de procesos/dimensiones que se muestra a continuación.

Procesos de negocio.	Dimensiones												
	Tiempo	Instancia	Materia	Estado	Centros	Involucrados	Medidas	Expediente	Persona	Respuestaimpugnacion	Tipopronunciamiento	Revisiones	Motivoacogida
Atención a la población	x	x	x						x				
Quejas y reclamaciones	x	x	x						x		x		
Impugnación	x	x	x			x				x			
Revisiones	x	x	x	x								x	
Visitas	x	x			x								

menores	x	x			x		x	x	x				x

Tabla 2. Matriz de procesos/dimensiones

1.1.3 Modelado Dimensional

Para la creación del modelo dimensional, se realizó la matriz descrita en el punto anterior, luego se definirán los procesos del negocio a los cuales se le realizara el mercado de datos y establecer el nivel de granularidad, posteriormente se describirán las tablas dimensiones y hechos, obteniendo con estas el modelo dimensional de alto nivel.

1.1.4 Elegir el proceso de negocio

El área de proceso a modelar seria del subsistema PDC los módulos: Atención a la Población, Revisiones Laborales. Revisiones Penales y Protección a Menores.

1.1.5 Establecer el nivel de granularidad

En este punto se muestra una tabla con la descripción de los datos asociados a las dimensiones, donde se explica el tipo de dato y/o los posibles valores que pueden tomar estos. A continuación se muestra la granularidad de estos en la siguiente tabla.

Nombre de la dimensión.	Nombre del dato	Tipo de dato	Valor nulo	Descripción.
tiempo	num_dia	Numeric(3,0)	no	Día en que se realiza una acción determinada.
	id_tiempo	Numeric(10,0)	No	Identificador de la tabla tiempo.
	num_mes	Numeric(3,0)	No	Mes en que se realiza una acción determinada.
	anno	Varchar(20)	no	Año en que se realiza una acción no determinada.

	num_trimestre	Numeric(3,0)	no	Trimestre en que se realiza una acción determinada.
	dia	Varchar(20)	no	Día en que se realiza una acción determinada.
	mes	Varchar(20)	no	Mes en que se realiza una acción determinada.
	trimestre	Varchar(20)	no	Trimestre en que se realiza una acción determinada.
	fecha	date	no	Fecha en que se crea el proceso.
materia	id_materia	Numeric(10)	no	Identificador de la materia
	nombre_materia	Varchar(255)	no	Nombre de la materia.
instancia	tipo de instancia	Varchar(255)	no	Instancia en que se analizan los datos (FGR, Provincial, Municipal).
	id_instancia	Numeric(19,0)	no	Identificador de la instancia.
involucrados	nombre_involucrados	Varchart(255)	no	Involucrados en una impugnación (promoventes, entidades infractoras).
	id_involucrados	Numeric(19,0)	no	Identificador de la tabla

estado	estado	varchart(255)	no	Estado del proceso.
	id_estado	Numeric(19,0)	no	Identificador de la tabla
medidas	tipo_medidas	Varchart(255)	no	Medidas aplicadas a los menores.
	id_medidas	Numeric(10,0)	no	Identificador de la tabla
expediente	causa_apertura	Varchart(255)	no	Causa de los expedientes abiertos a los menores.
	id_expediente	Numeric(19,0)	no	Identificador de la tabla
persona	id_persona	Numeric(19,0)	no	Identificador de la tabla
	genero	Varchart(255)	no	Sexo de la persona
	color_piel	Varchart (255)	no	Color de piel de la persona
	edad	numeric(3,0)	no	Edad del menor
motivoacogida	id_motivoacogida	Numeric (19,0)	no	Identificador de la tabla
	motivo_acogida	Varchart(255)	no	Motivo por el cual entra al centro de atención de menores
respuestaimpugnacion	id_respuestaimpugancion	Numeric (19,0)	no	Identificador de la tabla
	descripcion	Varchart(255)	no	Respuesta a la impugnación
tipopronunciamiento	id_tipopronunciamiento	Numeric (19,0)	no	Identificador de la tabla
	descripcion	Varchart(255)	no	Tipo de pronunciamiento

revisiones	id_revisiones	Numeric (19,0)	no	Identificador de la tabla
	tipo_revision	Varchart(255)	no	Tipo de revisión
centros	id_centro	Numeric (19,0)	no	Identificador de la tabla
	tipo_centro	Varchart(255)	no	Centros de atención a menores

Tabla 3. Nivel de granularidad

1.1.6 Identificación de las dimensiones

Las tablas dimensiones almacenan un conjunto de valores que están relacionados a una dimensión particular. Definen como están los datos organizados lógicamente y proveen el medio para analizar el contexto del negocio.

Para realizar estas tablas se tomará en cuenta las necesidades de información de los reportes para el subsistema PDC. La creación de la dimensión estará compuesta por los siguientes pasos:

1. Se elegirá un nombre que identifique la dimensión.
2. Se añadirá un campo que represente su clave principal.

Las dimensiones identificadas para modelar el Mercado de Datos se explicarán detalladamente a continuación para su mejor entendimiento:

Dimensión tiempo: esta dimensión situaría cada acción realizada en la fiscalía en el tiempo, se define esta dimensión con varias categorías o niveles para su mejor organización, estos son año, trimestre, mes y día.

Dimensión materia: define en que materia se realiza una acción determinada.

Dimensión instancia: definiría en que instancia de las fiscalías existentes de nuestro país se están obteniendo los datos, las cuales serían Fiscalía General de la Republica, Provincial y Municipal.

Dimensión estado: muestra el estado por los que pasan los procesos que se llevan a cabo en la fiscalía.

Dimensión Centros: serían los centros de atención a menores a los cuales se les realizan las visitas. Los cuales pueden ser Centros de Reeducación o EFI o Escuelas de Conducta.

Dimensión involucrados: personas o agentes que interviene en las impugnaciones, pueden ser promoventes o entidades infractoras.

Dimensión expediente: guarda los datos de los expedientes abiertos contra menores.

Dimensión medida: medidas aplicadas a los menores.

Dimensión persona: guarda los datos relacionados a las personas.

Dimensión respuestaimpugnacion: guarda la respuesta que se le da la impugnación. Un ejemplo de esta sería: resueltas fuera del término.

Dimensión tipopronunciamiento: declaración de la decisión de un juez a las reclamaciones existentes.

Dimensión revisiones: guarda los datos relacionados con las revisiones penales y las laborales.

Dimensión motivoacogida: motivos por el cual un menor ingresa a un centro de atención a menores.

1.1.7 Identificar las tablas de hechos y medidas

Una característica esencial de las tablas de hechos es que contienen datos numéricos (hechos) que se pueden resumir para proporcionar información sobre el historial de las operaciones de la organización. Además contiene, como claves externas, las claves primarias de las tablas de dimensiones asociadas a esta, así como los atributos de los registros de hechos.

A continuación se definirán las tablas de hechos a utilizar en el mercado de datos, a estas se le asignará un nombre, el cual representará el negocio enfocado:

Hecho Atención a la población: en esta tabla se recogerán todas las medidas relacionadas con la atención a la población en las diferentes instancias.

Hecho Visitas: aquí se guarda todo lo relacionado con las visitas a los menores por los diferentes centros de atención a ellos.

Hecho Quejas y reclamaciones: este hecho guarda todas las sumarizaciones relacionadas con los diferentes tipos de reclamaciones y quejas realizadas a la fiscalía.

Hecho Impugnación: este hecho guarda los aspectos relacionado con las impugnaciones y su estado.

Hecho revisiones: guarda todo lo relacionado con las revisiones penales, y laborales.

Hecho menores: guarda toda la información relacionada con los menores que han tenido problema de conducta o se encuentran en centros de acogida de menores o escuelas de conducta.

1.1.8 Modelo gráfico de alto nivel

Para concluir con el proceso dimensional inicial se realiza un gráfico denominado modelo dimensional de alto nivel, como ilustra la figura siguiente:

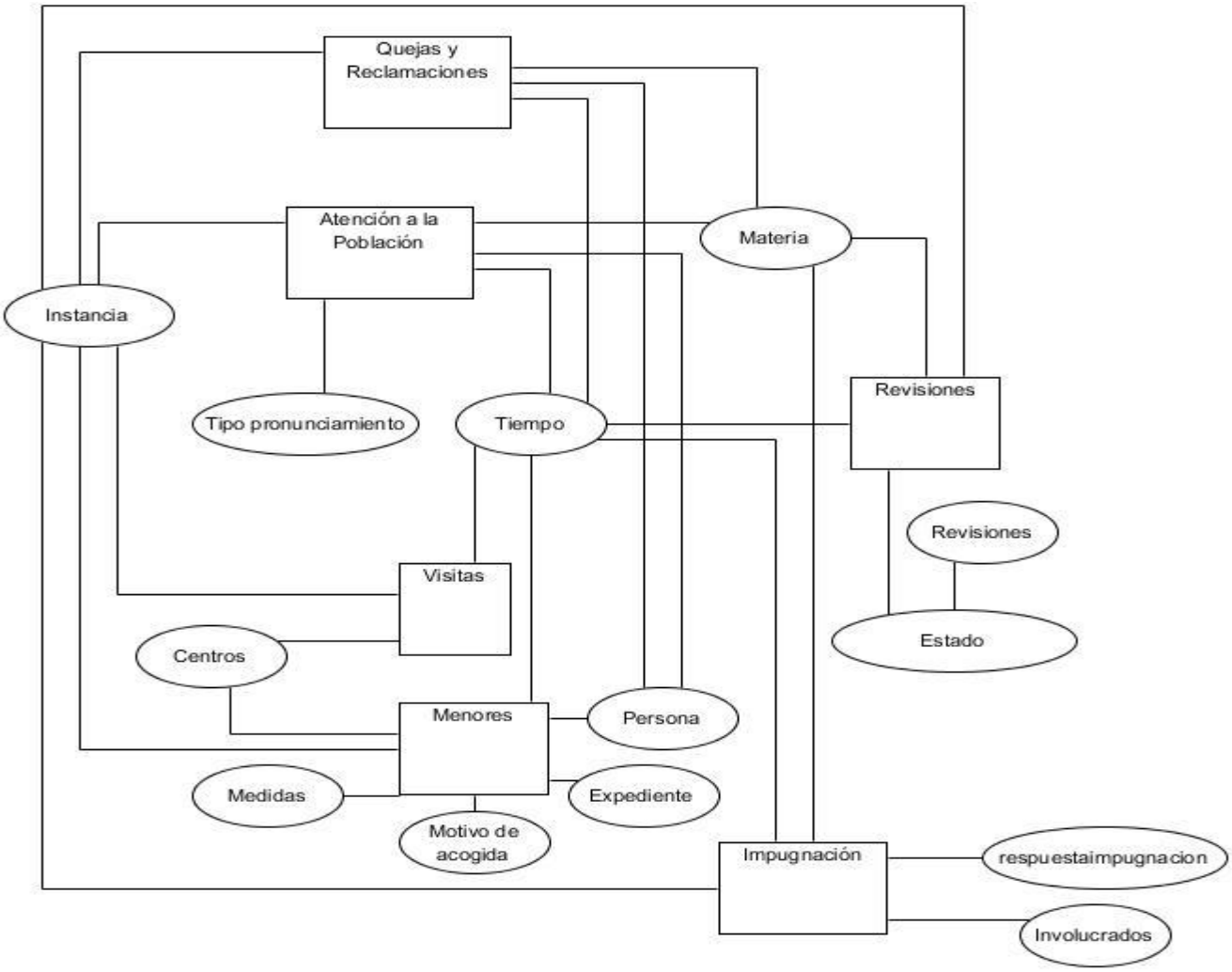


Ilustración 7 Modelo gráfico de alto nivel

1.1.9 Identificación de atributos de dimensiones y tablas de hechos

En la siguiente tabla se muestran todas las tablas dimensiones con sus respectivos atributos.

Nombre de la dimensión	Atributos
Dim_tiempo	id_tiempo, dia, mes, anno, trimestre
Dim_materia	nomb_materia, id_materia
Dim_centro	id_centro, tipo_centro
Dim_estado	id_estado, estado
Dim_instancia	id_intancia, tipo_instancia
Dim_involucrados	id_involucrados, nomb_involucrados
Dim_medidas	id_medidas, tipo_medidas
Dim_expediente	id_expediente, motivo_apertura
Dim_persona	id_persona, genero, edad, color_piel
Dim_respuestaimpugnacion	id_respuestaimpugnacion, descripcion
Dim_tipopronunciamiento	id_tipopronuncimaiento, descripcion
Dim_revisiones	id_revision, tipo_revision
Dim_motivoacogida	id_motivoacogida, motivo_acogida

Tabla 4. Atributos de las tablas dimensiones.

A continuación se muestran las tablas hechos con sus atributos:

Nombre de los hechos	Atributos
Hecho_atencion_poblacion	id_atencion_poblacion
Hecho_quejas_reclamaciones	id_quejas_reclamaciones
Hecho_impugnacion	id_impugnacion
Hecho_visitas	id_visitas
Hecho_revisiones	id_revisiones

Hecho_menores	id_menores

Tabla 5. Atributos de las tablas hechos.

1.2 Patrones de diseño de bases de datos para el MD

Los patrones constituyen una solución estándar para un problema común.

Los patrones de diseño de una Base de Datos permiten al usuario crear una BD más fortalecida ya que constituyen una guía que especifica cómo debe ser la misma. En la actualidad las bases de datos suelen ser muy grandes y a veces el trabajo con los patrones de diseño hacen que el diseño sea más fácil y que se asegure un resultado satisfactorio.

Patrón de llaves subrogadas

Este patrón es muy utilizado pues facilita la interacción con la BD en un futuro. El mismo plantea que se genere una llave primara única para cada entidad, en vez de usar un atributo identificador en el contexto dado.

Esto permite que las tablas sean más fáciles de consultar a partir del identificador, pues todos tienen el mismo tipo en cada una de las tablas. Por ello cada tabla incluyendo las tablas hechos tendrá una llave única que la identifique.

1.3 Estandarización del código

Se conoce por estandarización al proceso mediante el cual se realiza una actividad de manera previamente establecida. El término estandarización proviene del término estándar, aquel que refiere a un modo o método establecido, aceptado y normalmente seguido para realizar determinado tipo de actividades o funciones. Un estándar es un parámetro más o menos esperable para ciertas circunstancias o espacios y es aquello que debe ser seguido en caso de recurrir a algunos tipos de acción. El término de estandarización tiene como connotación principal la idea de seguir entonces el proceso estándar a través del cual se tiene que actuar o proceder.

A continuación se explican los estándares de código a seguir para la realización del mercado de datos del subsistema PDC del SIGEF:

- Los nombres de las entidades y atributos serán con minúscula, separando por "_" cuando exista composición de palabras.

- Las palabras que tenga ñ esta serán tratadas como "nn".

- Las tablas hechos comenzarán con la palabra “hecho”.
- Las tablas dimensiones comenzarán con “dim”.
- Los identificadores de las tablas comenzarán su nombre con “id”.
- Las palabras que lleven tilde se pondrán sin estas en las tablas.

1.4 Tipo de modelo lógico del mercado de datos

Es importante señalar que para la determinación del modelo se tuvo en cuenta los requerimientos y necesidades de los usuarios ya que es de suma importancia definir cuál esquema se empleará para el diseño de la estructura del mercado de datos debido a que esta selección afectará considerablemente la elaboración del modelo lógico.

La variante de modelamiento que se propone utilizar en el desarrollo del mercado de datos para el subsistema PDC es **Esquema Constelación,** debido a las necesidades de los usuarios. Se selecciona este esquema porque existirán tablas de dimensiones que se relacionarán con más de una tabla de hechos.

1.5 Arquitectura del mercado de datos

Teniendo en cuenta las características generales de un mercado de datos se definen los componentes que intervienen en su arquitectura o ambiente. En el siguiente gráfico se muestra la estructura que tendrá el almacenamiento de datos.

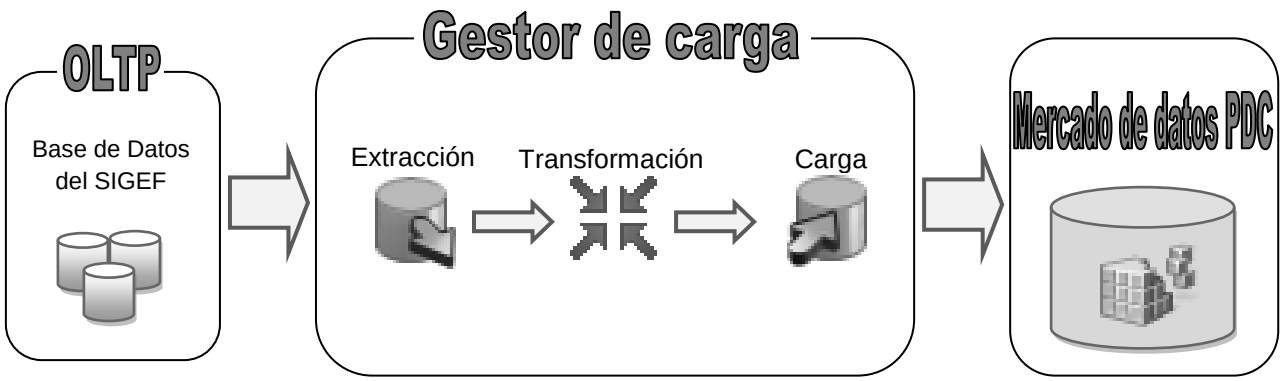


Ilustración 8 Arquitectura propuesta para el mercado de datos.

El ambiente está formado por diversos elementos que interactúan entre sí y que cumplen una función específica dentro del sistema. Básicamente la forma de operar del esquema superior se resume de la siguiente manera:

- Los datos son extraídos de las aplicaciones, bases de datos, archivos, etc. Esta información reside generalmente en diferentes tipos de sistemas, orígenes y arquitecturas, también tienen formatos muy variados.
- Los datos son transformados, limpiados e integrados para luego ser cargados en el mercado de datos.
- La información del mercado de datos se estructura en cubos multidimensionales, aunque también se pueden utilizar otros tipos de estructuras para representar la información del mercado de datos.

Como se había visto anteriormente los OLTP representan toda la información transaccional que genera la institución en su accionar diario, además de todas las fuentes de datos externas que puedan existir. En este caso la información de la que se habla se encuentra en las bases de datos que dispone el sistema SIGEF.

En el gestor de carga se utilizará un sistema que es capaz de extraer los datos, manipularlos, transformarlos e integrarlos, este sistema es comúnmente conocido como sistemas ETL. Este proceso es realizado con la herramienta Pentaho Data Integration. A continuación se explica en síntesis el accionar del proceso de extracción, transformación y carga:

- **Extracción:** extrae los datos relevantes, de la fuente de datos.
- **Transformación:** convierte datos desde su formato OLTP al apropiado del mercado de datos, ejecutando funciones tales como convertir una variedad de formatos de fechas hacia un formato estándar y renombrando campos desde nombres técnicos no significativos hacia nombres significativos que el usuario final entenderá.
- **Carga:** los datos de la fuente normalmente son extraídos y cargados a la base de datos del mercado de datos.

1.6 Modelado conceptual de los datos

Después de realizar el análisis de las necesidades de información y las fuentes de datos se llegó al siguiente modelado conceptual de la BD:

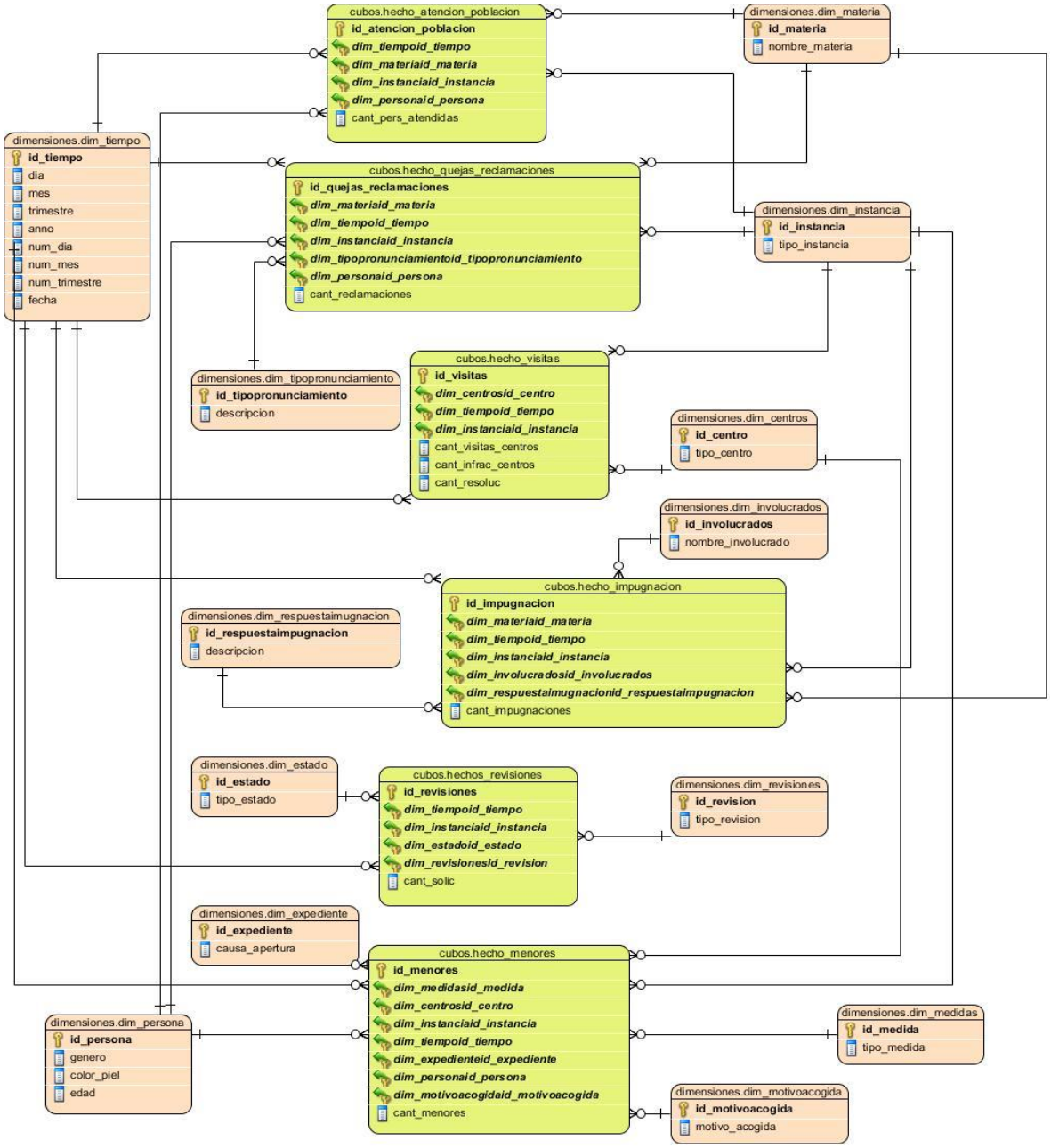


Ilustración 9 Modelo conceptual de los datos

Conclusiones del capítulo

En el desarrollo del presente capítulo se brinda una breve descripción del negocio para lograr una mejor comprensión a la hora de analizar y diseñar el mercado de datos. También se realizó un detallado análisis de las diferentes necesidades de información que existen. Además se estudió el estado actual de las fuentes de datos con que se cuenta actualmente. Otro de los resultados es establecer las correspondencias entre el resultado del análisis de las necesidades de información con las fuentes de datos, dando paso al posterior diseño del mercado de datos.

Además se detallaron todos los aspectos relacionados al diseño del mercado de datos. Primeramente se explican los elementos sobre la arquitectura que tendrá la solución propuesta. Se muestran un conjunto de aspectos que conforman los estándares para el diseño de los almacenes de datos. Se realizó un estudio de los patrones de diseño de BD y se concluye que el patrón a utilizar sería el patrón llave subrogada. El principal resultado en el desarrollo del capítulo es la identificación de las dimensiones y hechos, además de las uniones establecidas entre las mismas, que forman el diseño del mercado de datos.

Capítulo 3:Implementación.

3.1 Introducción.

Una vez diseñado el mercado de datos, se realiza el proceso de ETL de los datos para así limpiar estos y que los mismos puedan ser ingresados al mercado de datos. En este capítulo se realiza el diseño de los procesos de ETL para el mercado de datos del SIGEF. Se implementan los cubos multidimensionales que dan como resultado los reportes que responden las necesidades de información del usuario.

3.2 Diseño físico de la base de datos.

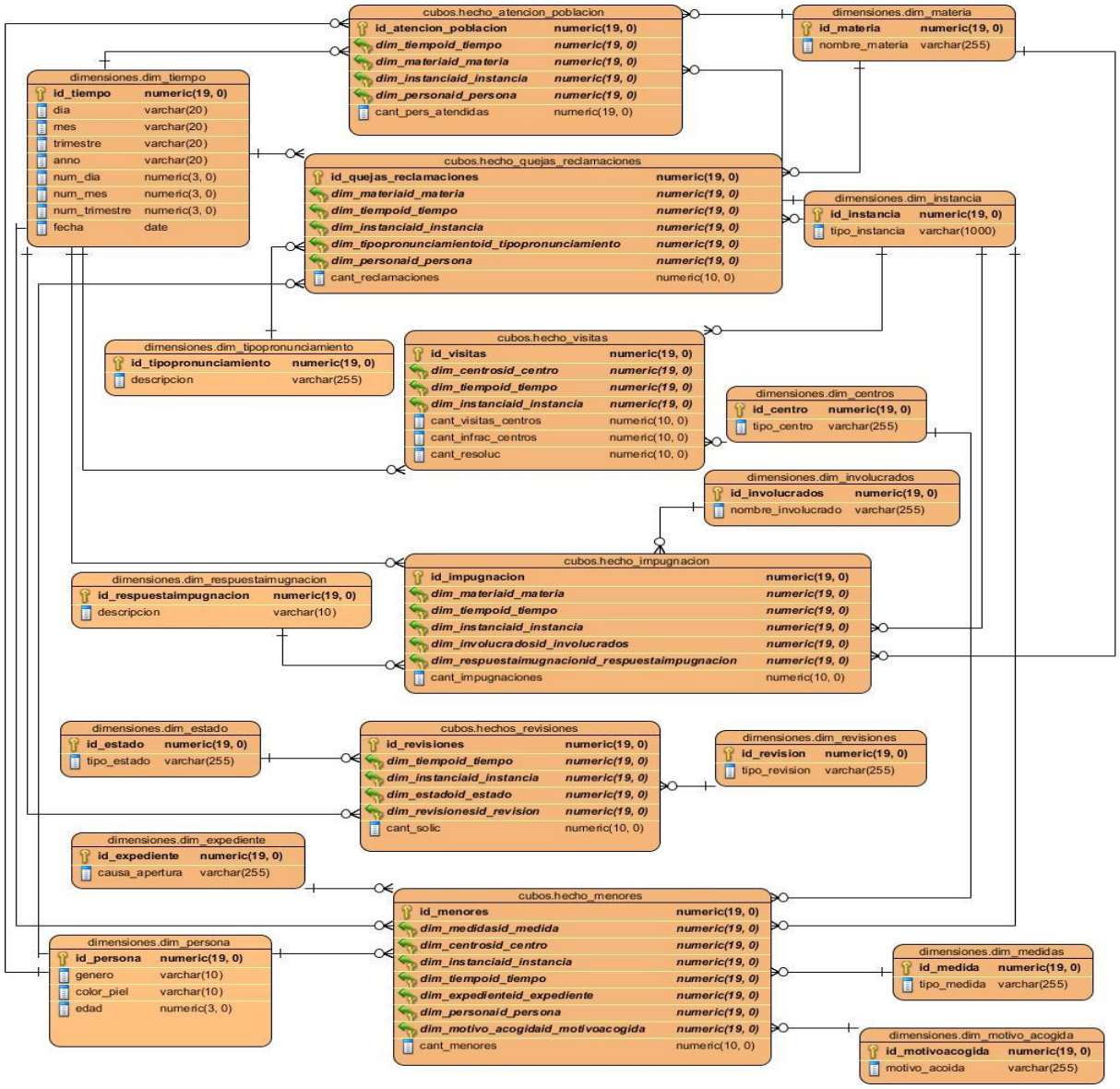


Ilustración 10 Diseño físico de la base de datos

3.3 Implementar el modelo dimensional detallado

Con el objetivo de lograr una mejor organización en el trabajo, evitar las pérdidas de datos y posibilitar una buena ejecución de los pasos siguientes, se realiza un mapeo de datos, indicando de donde se extrajeron los datos en la BD del SIGEF y hacia donde se trasladaron en el mercado de datos PDC; detallando los campos fuentes y destino, así como el tipo de dato, el mismo se presentan en el anexo 16 “Mapa Lógico de Datos.doc” se muestra detalles del mapeo de datos realizado.

3.4 Prueba del modelo

Se le realizan pruebas al modelo haciendo preguntas para ver si puede responder a estas, un ejemplo de esta seria:

¿Cuántas visitas se han realizado a centros de atención a menores, por una instancia en un periodo determinado?

¿Cuántas visitas se han realizado a un centro determinado?

¿Cuántas infracciones se han detectado en un centro determinado?

Como se puede observar en la Ilustración 12 la tabla hecho_visitas tiene relación con las dimensiones: tiempo, instancia, centros; además contiene las medidas cant_visitas y cant_infrac_centros, por lo que el modelo está apto para responder a estas preguntas.

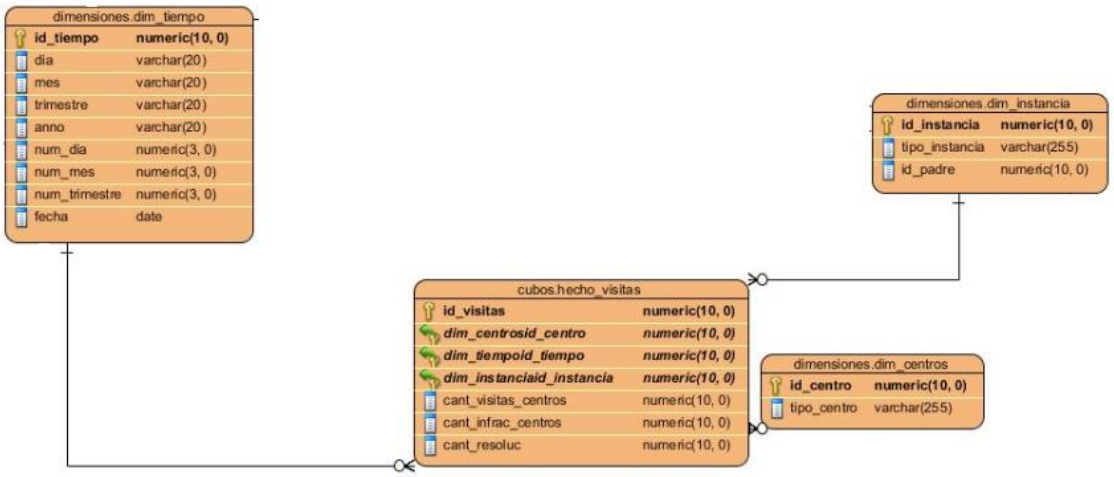


Ilustración 11 Fragmento del modelo conceptual.

3.5 Diseño de los sistemas de extracción, transformación y carga de los datos.

Antes de comenzar el diseño de las transformaciones de los datos fue necesario realizar una configuración de la conexión a la base de datos fuente. La cual se puede observar en el Anexo 1.

El próximo paso sería el diseño de la transformación donde se convierten los datos inconsistentes en un conjunto de datos compatibles para que puedan ser cargados en el mercado de datos. Entre los pasos más comunes que se realizaron se encuentran: entrada de tabla, mapeo de valores, búsquedas en base de datos, insertar/actualizar y actualizar datos. En los Anexos del 1 al 5 se puede observar los componentes. El diseño de las transformaciones se muestra en el Anexo 6 mediante imágenes. A continuación se muestra el diseño de la transformación para la dimensión centros (Ilustración 12) y el hecho visitas (Ilustración 13).

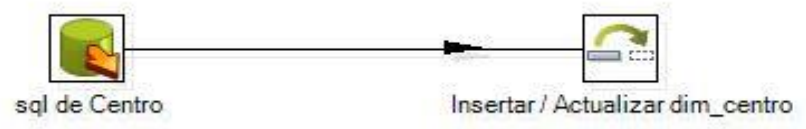


Ilustración 12 Transformación para la dimensión "dim_centro"

Ilustración 13 Transformación para el hecho "hecho_visitas"

Basándose en las necesidades de información de la fiscalía para el subsistema PDC, se explora la fuente de datos a disposición, y se extrae la información que se considere relevante. Para la realización de este proceso se crearon un conjunto de consultas SQL (Anexo 7) que dan cumplimiento a este objetivo. Las mismas se incorporan al componente de entrada de Kettle.

Por último se cargan los datos hasta el mercado de datos y se le realiza actualizaciones o mantenimientos periódicos, esto se realiza a partir de las transformaciones previamente diseñadas. Para dar cumplimiento a la misma se diseñó un trabajo para cada tabla de hecho, donde primero se cargarán los datos de las dimensiones asociadas a este y luego se cargan los de la tabla de hecho. También fue necesario crear un trabajo general denominado "general" donde se realiza una carga inicial de todos los datos. Esta carga inicial se refiere precisamente a la primera carga

de datos que se realiza al mercado de datos, por lo general esta tarea consume bastante tiempo, ya que se deben de insertar todos los registros que han sido generados. Por otra parte, los restantes trabajos tienen la función de realizar los mantenimientos periódicos que mueven pequeños volúmenes de datos, y su frecuencia está dada por el gránulo y los requerimientos de los usuarios. Estos trabajos se diseñaron haciendo uso de los componentes que brinda la herramienta utilizada (Anexos 8, 9 y 10). Entre los pasos más comunes que se realizaron se encuentran: inicio de un trabajo, ejecutar una transformación y fin. El detalle de los trabajos diseñados se muestra en el Anexo 11 a través de imágenes.

3.6 Automatización del sistema ETL

Teniendo en cuenta la frecuencia de actualización de los datos y las necesidades de información del cliente en el sistema SIGEF, se propone que los procesos de extracción, transformación y carga se realicen mensualmente y de forma automática. El objetivo de esta automatización es añadir al mercado de datos aquellos datos nuevos que fueron generados después de la última actualización.

Esta actualización se realiza con Kitchen que es un programa que puede ejecutar trabajos diseñados con Spoon y almacenados como XML ó en el repositorio de base de datos.

Como el servidor instalado en las fiscalías es Ubuntu server 12.04 se utiliza el comando “crontab –e” para programar la ejecución de los trabajos previamente diseñados.

Después de ejecutado este comando se procede a la programación de la tarea. A continuación se explican los componentes de esta para un mejor entendimiento:

- El minuto de la hora, 0-59
- La hora del día, 0-23
- El día del mes, 1-31
- El mes del año, 1-12
- El día de la semana, 0-6, 0=Domingo

Cuando es usado un asterisco ("*") en lugar de un número, significa cualquier número posible para esa posición.

Actualizar_trabajo.sh se refiere a un script creado que manda a Kitchen a ejecutar los trabajos previamente diseñados.

#

Ejecuta la actualización de las dimensiones del almacén de datos

#

1,0,1, * * * /dirección del directorio/actualizar_trabajos.sh

#

La tarea quedo programada para todos los meses el dia primero a las 12:01 am.

3.7 Diseño de las dimensiones y cubo de información

Una vez terminado el diseño de las transformaciones, trabajos y la carga los datos al mercado de datos, se procede a establecer una conexión al mercado de datos (Anexo 12), posteriormente se procede a realizar diseño de la dimensiones y los cubos de información en la herramienta Schema-Worbench.

Se comienza diseñando los cubos multidimensionales y dentro de estos las medidas y dimensiones definiendo para esta última las jerarquías, niveles y propiedades. Las dimensiones no se crean en un cubo específicamente, las misma se crean fuera, para que otros cubos puedan utilizarlos si llega a ser necesario.

A continuación se muestra la creación multidimensional en el Schema--Worbench.

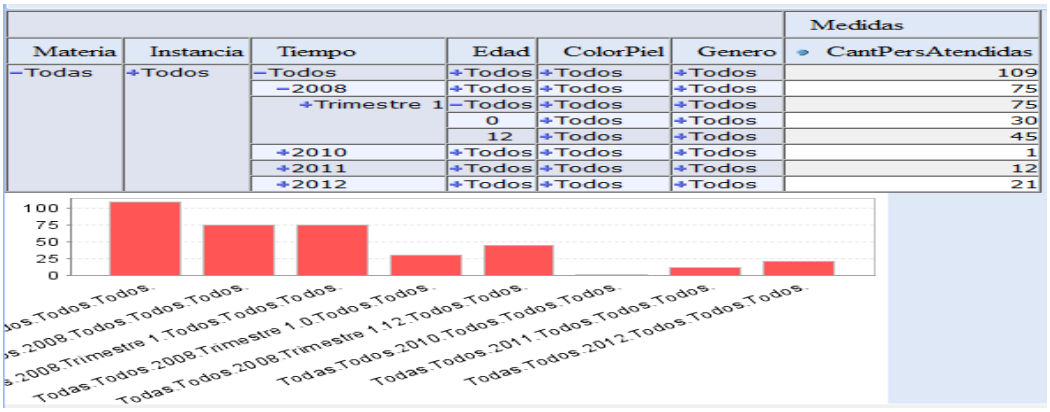
1. Se crean las dimensiones. (Anexo 13)

2. Se crean los cubos. (Anexo 14)

Una vez diseñado las dimensiones y cubos se representa la información referente a las necesidades del cliente, la cual se muestra mediante la ejecución de las consultas MDX a través del motor OLAP Mondrian. En el Anexo 15 se muestra un ejemplo y la consulta MDX que le corresponde al reporte atención a la población.

3.8 Visualización de los resultados:

Las consultas MDX son las encargadas de decir que dimensiones y medidas se van a representar en el navegador, a cada cubo se le realiza una consulta MDX para especificar que dimensiones y medidas se deben mostrar. A continuación se muestra como se visualizan estos datos en el navegador para el cubo Atención a la población.



						Medidas
Materia	Instancia	Tiempo	Edad	ColorPiel	Genero	CantPersAtendidas
−Todas	+Todos	−Todos	+Todos	+Todos	+Todos	109
		−2008	+Todos	+Todos	+Todos	75
		+Trimestre 1	−Todos	+Todos	+Todos	75
			0	+Todos	+Todos	30
			12	+Todos	+Todos	45
		+2010	+Todos	+Todos	+Todos	1
		+2011	+Todos	+Todos	+Todos	12
		+2012	+Todos	+Todos	+Todos	21

Ilustración 14 Representación en el navegador del cubo Atención a la Población.

Conclusiones del capítulo:

En este capítulo se realizó el mapeo de datos para definir las tabla origen de donde se extraería la información y las tablas destino hacia donde se dirige la misma, se explicaron los pasos realizados para diseñar el proceso de extracción, transformación y carga de los datos. Se explicó cómo se realiza el diseño del modelo multidimensional y cómo este es mostrado en reportes que responden a las necesidades de información del cliente.

.

Capítulo 4: Pruebas de validación

4.1 Introducción:

Dentro del desarrollo de los almacenes de datos la realización de las pruebas es un paso importante para garantizar el éxito de la solución informática. El mecanismo que se seguirá para ello será la realización de pruebas de volumen y carga, además de pruebas de carga y estrés, las cuales validarán la utilización del mercado de datos. En este punto se analizan los rendimientos del sistema que se ha construido, al dar respuesta a distintos pedidos de información accediendo a la base de datos que se encuentra en el servidor PostgreSQL.

4.2 Pruebas de Volumen y Carga

La prueba de volumen y carga consiste en insertar una gran cantidad de datos al mercado de datos, para determinar si se alcanzan límites que hagan fallar el software. También identifica la carga máxima continua o volumen que el elemento de prueba puede manejar por un período de tiempo dado. La prueba de volumen permite verificar que la aplicación funciona adecuadamente con el máximo tamaño esperado de base de datos.

Para la realización de la prueba de carga se utilizó la herramienta Data Generator para PostgreSQL, la cual permitió llenar el mercado de datos con una cantidad determinada de datos. Se configuró esta generación con valores arbitrarios, pero coincidentes en cuanto a su tipo y volumen.

La cantidad de valores determinados para poblar el mercado de datos fue 20 000 tuplas por tabla. Para la definición de esta cantidad se tuvo en cuenta que anualmente se espera que ocurra un aproximado de 200000 procesos entre todas las fiscalías del país, como el mercado de datos solo responde a las necesidades de información del subsistema PDC, de los módulos Atención a la población, Menores, Revisiones Laborales y Revisiones Penales, solo se guardan en el los datos de los procesos que están relacionados a las necesidades de información de los mismos. Se estiman que en los procesos que abarca el mercado de datos se produzcan un total de 76000 tuplas anuales, que en un periodo de 5 años (tiempo de vida del mercado de datos) alcanzaría las 380000 tuplas. Transcurrido los 5 años se procede a realizar una salva de los datos, se limpia el mercado y se procede a comenzar a cargarlo nuevamente. A

.

continuación se muestra el resultado de la prueba de volumen realizada en la herramienta.

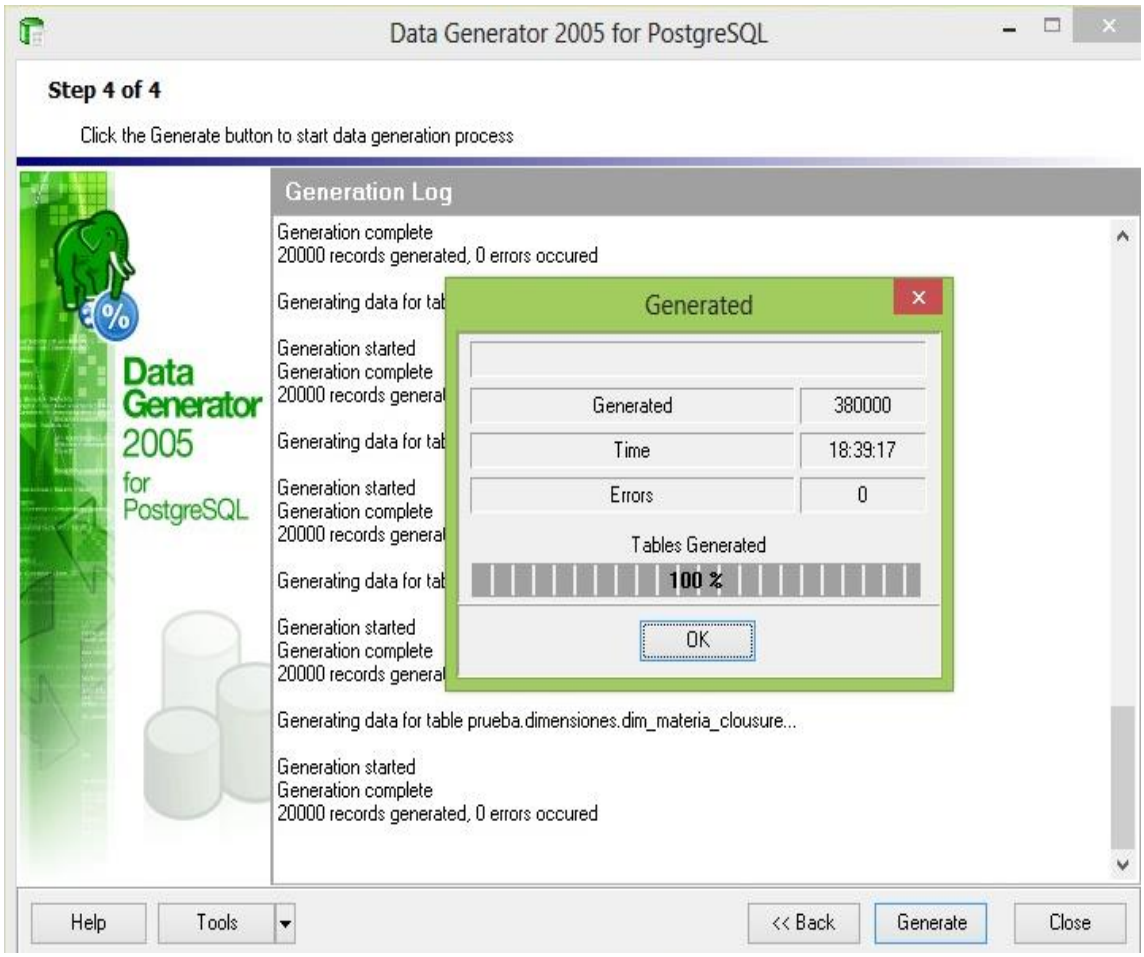


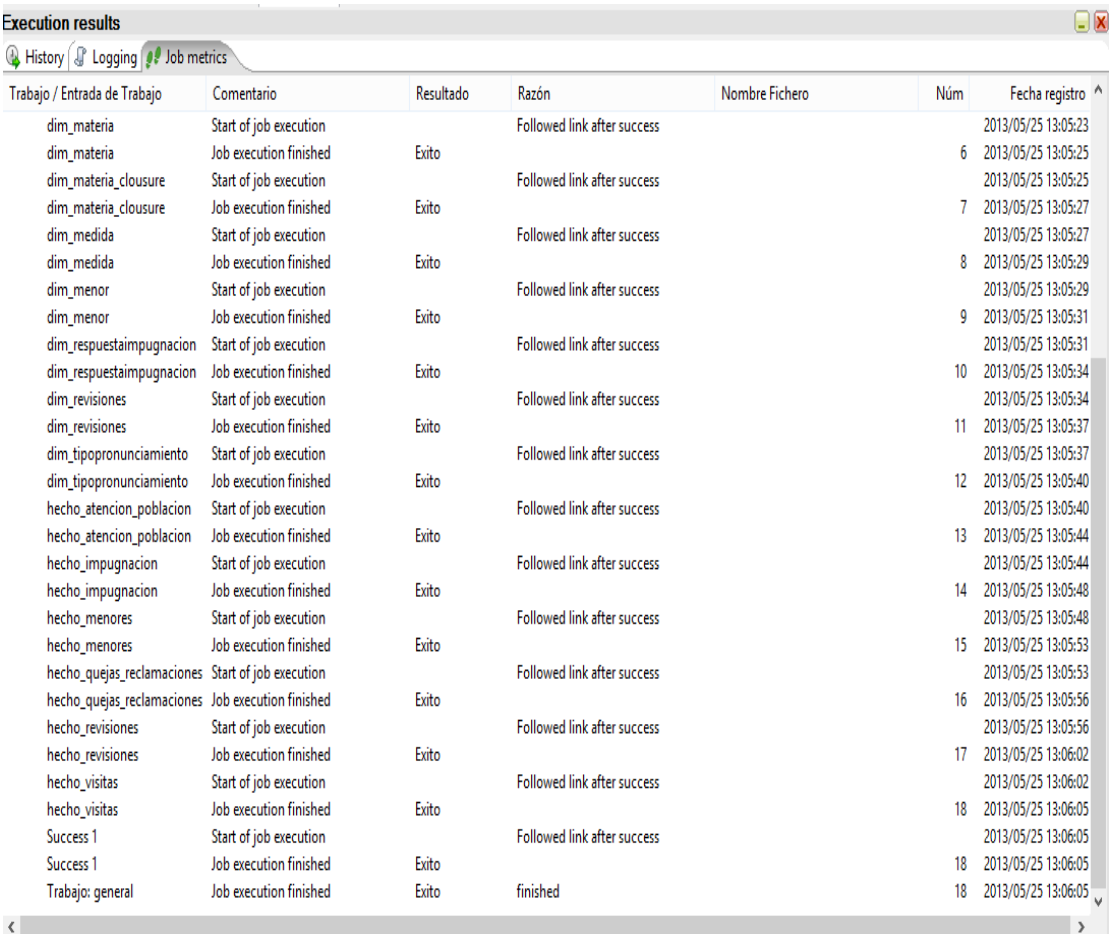
Ilustración 15 Resultado de la prueba de Carga en Data Generator.

Al introducir los datos no se presentaron problemas de límite de capacidad, ni se detectaron desbordamientos de columnas, atributos, tipos de datos, ni peticiones excesivas de memoria.

Las llaves autogeneradas no se salieron del rango especificado, ni se detectaron problemas con los tipos de datos definidos en el paso de diseño. Lo anteriormente planteado garantiza que el SGBD utilizado y el diseño implementado portan completamente el almacenamiento de los niveles de información requeridos para la puesta en producción del mercado de datos para PDC.

.

Otro elemento utilizado como prueba de carga fue la ejecución del trabajo general que es el encargado de cargar todos los datos de la base de datos relacional al mercado de datos, esta ejecución corrió con éxito a continuación se muestra dicha ejecución:


Execution results

History | Logging | Job metrics

Trabajo / Entrada de Trabajo	Comentario	Resultado	Razón	Nombre Fichero	Núm	Fecha registro
dim_materia	Start of job execution		Followed link after success			2013/05/25 13:05:23
dim_materia	Job execution finished	Exito			6	2013/05/25 13:05:25
dim_materia_clousure	Start of job execution		Followed link after success			2013/05/25 13:05:25
dim_materia_clousure	Job execution finished	Exito			7	2013/05/25 13:05:27
dim_medida	Start of job execution		Followed link after success			2013/05/25 13:05:27
dim_medida	Job execution finished	Exito			8	2013/05/25 13:05:29
dim_menor	Start of job execution		Followed link after success			2013/05/25 13:05:29
dim_menor	Job execution finished	Exito			9	2013/05/25 13:05:31
dim_respuestaimpugnacion	Start of job execution		Followed link after success			2013/05/25 13:05:31
dim_respuestaimpugnacion	Job execution finished	Exito			10	2013/05/25 13:05:34
dim_revisiones	Start of job execution		Followed link after success			2013/05/25 13:05:34
dim_revisiones	Job execution finished	Exito			11	2013/05/25 13:05:37
dim_tipopronunciamiento	Start of job execution		Followed link after success			2013/05/25 13:05:37
dim_tipopronunciamiento	Job execution finished	Exito			12	2013/05/25 13:05:40
hecho_atencion_poblacion	Start of job execution		Followed link after success			2013/05/25 13:05:40
hecho_atencion_poblacion	Job execution finished	Exito			13	2013/05/25 13:05:44
hecho_impugnacion	Start of job execution		Followed link after success			2013/05/25 13:05:44
hecho_impugnacion	Job execution finished	Exito			14	2013/05/25 13:05:48
hecho_menores	Start of job execution		Followed link after success			2013/05/25 13:05:48
hecho_menores	Job execution finished	Exito			15	2013/05/25 13:05:53
hecho_quejas_reclamaciones	Start of job execution		Followed link after success			2013/05/25 13:05:53
hecho_quejas_reclamaciones	Job execution finished	Exito			16	2013/05/25 13:05:56
hecho_revisiones	Start of job execution		Followed link after success			2013/05/25 13:05:56
hecho_revisiones	Job execution finished	Exito			17	2013/05/25 13:06:02
hecho_visitas	Start of job execution		Followed link after success			2013/05/25 13:06:02
hecho_visitas	Job execution finished	Exito			18	2013/05/25 13:06:05
Success 1	Start of job execution		Followed link after success			2013/05/25 13:06:05
Success 1	Job execution finished	Exito			18	2013/05/25 13:06:05
Trabajo: general	Job execution finished	Exito	finished		18	2013/05/25 13:06:05

Ilustración 16 Prueba realizada a la ejecución del trabajo general

4.3 Pruebas de carga y estrés

Para realizar las pruebas de carga y estrés se utiliza la herramienta JMeter en su versión 2.4 por la facilidad de su uso y las funcionalidades que brinda. Esta herramienta posee dos tipos de generación de carga, indirecta, es decir, a través de una aplicación y directa que basa fundamentalmente su utilización en la ejecución de consultas grabadas en la traza o log del servidor de base de datos. La que se va a utilizar para las pruebas del mercado de datos es la directa.

A continuación se muestra la cantidad de usuarios potenciales que tendrá el mercado de datos en los distintos niveles de las FGR:

.

Instancia	Cantidad de usuarios
Fiscalía general de la república	25
Fiscalías provinciales	12

Tabla 6. Cantidad de usuarios concurrentes en las diferentes instancias de las fiscalías (Machado, 2012).

Es preciso definir un rango de tiempo aceptable para las consultas, el cual queda definido de la siguiente forma:

Tipo de respuesta	Tiempo
Aceptable	Menor a los 5 segundos
Superior	Mayor a los 5 segundos

Tabla 7. Tiempo de respuestas de las consultas.

Fue necesario diseñar 2 pruebas para ver el tiempo de respuesta del mercado de datos. Se definieron la cantidad de hilos concurrentes en consecuencia con un estimado 2 veces mayor a la cantidad de usuarios que se van a conectar en una situación extrema en las diversas instancias de las fiscalías. La programación de estas pruebas se muestra a continuación:

Pruebas	Instancia	Cantidad de usuarios reales	Cantidad de hilos a probar
Prueba 1	Fiscalías provinciales	12	25
Prueba 2	Fiscalía general de la república	25	50

Tabla 8. Pruebas diseñadas.

Las pruebas consisten en realizar una consulta y ver los tiempos de respuestas que tiene para varios usuarios concurrentemente, la consulta utilizada se muestra a continuación:

```
SELECT
  cubos.hecho_menores.id_menores
FROM
  cubos.hecho_menores
  INNER JOIN dimensiones.dim_medidas
  ON (cubos.hecho_menores.dim_medidasid_medida = dimensiones.dim_medidas.id_medida)
  INNER JOIN dimensiones.dim_menor
  ON (cubos.hecho_menores.dim_menorid_menor = dimensiones.dim_menor.id_menor)
  INNER JOIN dimensiones.dim_instancia
  ON (cubos.hecho_menores.dim_instanciaid_instancia = dimensiones.dim_instancia.id_instancia)
  INNER JOIN dimensiones.dim_tiempo
  ON (cubos.hecho_menores.dim_tiempoid_tiempo = dimensiones.dim_tiempo.id_tiempo)
  INNER JOIN dimensiones.dim_centros
  ON (cubos.hecho_menores.dim_centrosid_centro = dimensiones.dim_centros.id_centro)
  INNER JOIN dimensiones.dim_expediente
  ON (cubos.hecho_menores.dim_expedienteid_expediente = dimensiones.dim_expediente.id_expediente)
```

Ilustración 17 Consulta usada para la prueba en JMeter.

En el informe agregado del JMeter se muestran una serie de datos los cuales exponen el estado en el que se encuentra en software, los cuales son los siguientes:

- Muestras: Cantidad de páginas (Hilos) que simulan la cantidad de usuarios que están interactuando con el sistema desde la misma URL.
- Media: Media de tiempo total que demoraron las petición en cargarse.
- Mediana: Tiempo promedio que han tardado en cargarse las páginas.
- Min: Tiempo mínimo que ha demorado en cargarse una página.
- Max: Tiempo Máximo que ha tardado en cargarse una página.
- Línea 90 %: Tiempo máximo en que corrieron el 90 por ciento de las peticiones reales, o sea, el tiempo más probable que se puede demorar una petición.
- Rendimiento: Cantidad de solicitudes por segundo.
- %Error: Por ciento de error de las páginas que no se llegaron a cargar de manera satisfactoria.
- Kb/Seg: Velocidad de carga de las páginas. (Almenares, 2008)

A continuación se muestran los resultados que arrojaron las pruebas:

# Muestras	Media	Mediana	Linea de 90%	Mín	Máx	% Error	Rendimiento	Kb/sec
25	4	4	5	2	11	0,00%	26,3/sec	,0
25	4	4	5	2	11	0,00%	26,3/sec	,0

Ilustración 18 Resultado para la prueba 1.

.

.

Media	Mediana	Linea de 90%	Mín	Máx	% Error	Rendimiento	Kb/sec
4193	4730	5177	13	5538	0,00%	8,9/sec	,0
4193	4730	5177	13	5538	0,00%	8,9/sec	,0

Ilustración 19 Resultado para la prueba 2.

Se puede concluir que los tiempos de respuesta son aceptables ya que ninguno excede a los 5 segundos.

Conclusiones del capítulo:

En este capítulo se realizaron un conjunto de pruebas que validan la solución propuesta, demostrando que el mercado de datos está apto para recibir el volumen de datos, que se espera surjan de los procesos tenidos en cuenta en la investigación y que es capaz de mantener el máximo de conexiones estimadas de manera concurrentemente. Además se midió el tiempo de ejecución de las consultas de acceso a los datos y las mismas cumplen con los umbrales de aceptación definidos por los clientes.

.

Conclusiones:

El estudio de los conceptos, características, herramientas y tecnologías, facilitó la selección de la metodología propuesta por Kimball por su característica de desarrollo incremental, y fácil comprensión, además de la selección de la suite de Pentaho para el desarrollo de la solución.

Se analizaron las diferentes necesidades de información que posee el cliente, para identificar las dimensiones y hechos, señalando como dimensiones importantes la dimensión tiempo y la dimensión instancia, además se identifican las uniones establecidas entre estos elementos, que forman el diseño del mercado de datos.

Se analizó el estado actual de las fuentes de datos, que arrojó como resultado las tablas importantes de estas para el mercado de datos, el cual facilitó la realización del proceso de extracción, transformación y carga, además de la creación del modelo multidimensional.

Se realizaron un conjunto de pruebas que validan la solución propuesta, las mismas prueban al sistema en situaciones extremas de concurrencia y carga de datos, dando como resultado que el mercado de datos está apto para responder a las necesidades de información del cliente.

El mercado de datos para los subprocesos revisiones laborales, revisiones penales, atención a la población y menores, permitiendo agilizar el análisis de los datos generados por el SIGEF II, favoreciendo el proceso de toma de decisiones en la institución.

.

Recomendaciones

Con el objetivo de mejorar la solución planteada se proponen las siguientes recomendaciones:

- Continuar el desarrollo de la solución, implementando mercados de datos para los restantes subsistemas del SIGEF.
- Tener en cuenta las nuevas necesidades de información que presenten los usuarios.
- Utilizar la metodología y las herramientas seleccionadas en esta investigación para el desarrollo de futuros mercados de datos en el SIGEF.

Bibliografía

Portada sobre la plataforma Pentaho Open Source Business Intelligence . 2006-2011. Portada sobre la plataforma Pentaho Open Source Business Intelligence. [En línea] 2006-2011. [Citado el: 03 de 12 de 2012.] http://pentaho.almacen-datos.com/.

2xMil Soluciones. 2012. Desarrollo de informes con Crystal Reports.NET(documento). *2xMil Soluciones informáticas.* [En línea] 2012. [Citado el: 03 de 12 de 2012.] http://www.2xmil.es/pdf/CURSO_CR.pdf.

AsiConcultant. AsiConcultant. [En línea] [Citado el: 06 de diciembre de 2012.] http://www.asiconsultant.com/soluciones_integracion.php?pagp_id=48&orden_id=17.

Bernabeo, Dario Ricardo. 2009. *DATA WAREHOUSING: Investigación y Sistematización de Conceptos - HEFESTO.* Argentina : s.n., 2009.

Ciberaula. 2010. Ciberaula Linux. [En línea] 2010. [Citado el: 07 de 12 de 2012.] http://linux.ciberaula.com/articulo/linux_apache_intro.

Dario, Ing. Bernabeu R. 2009. Datawarehouse manager. *Dataprix.* [En línea] 6 de mayo de 2009. [Citado el: 3 de abril de 2013.] http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager.

Díaz, Josep Curto. 2010. *Introducción al Business Intelligence.* Barcelona : UOC, 2010. ISBN:978-84-9788-886-8.

ETL-Tools.Info. 2006-2012. ETL-Tools.Info. [En línea] 2006-2012. [Citado el: 23 de 1 de 2013.] http://etl-tools.info/es/bi/almacendedatos_esquema-estrella.htm.

Gravitar. 2012. Gravitar. [En línea] 2012. [Citado el: 03 de 12 de 2012.] http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/.

guia-ubuntu. 2008. guia-ubuntu. [En línea] 10 de marzo de 2008. [Citado el: 03 de 12 de 2012.] http://www.guia-ubuntu.org/index.php?title=PgAdmin_III.

Informática_hoy. 2007-2012. Informática Hoy. [En línea] 2007-2012. [Citado el: 23 de 1 de 2013.] www.informatica-hoy.com.ar.

Inmon, W. H. 2002. *Building the Data Warehouse Third Edition.* NEW YORK, EEUU : John Wiley & Sons, Inc., 2002. ISBN: 0-471-08130-2.

Jmeter, Apache. 2013. Apache . [En línea] 2013. www.apache.org.

—. 2013. The Apache Software Fundation. [En línea] 22 de 01 de 2013. www.apache.org.

Kimball, Ralph, y otros. 2008. *The Data Warehouse Lifecycle Toolkit.* s.l. : Wiley, 2008. ISBN-10: 0470149779.

Lorenzo., Orestes Rodríguez. 2010. *Almacén de datos para los subsistemas de Reclutamiento y Potencial Humano.* La Habana : s.n., 2010.

Martinez, Rafael. 2009-2012. postgresql-es. [En línea] 2009-2012. [Citado el: 06 de 12 de 2012.] http://www.postgresql.org.es/sobre_postgresql.

Martinez, Yaneisy Contreras y González, Alicia Guilarte. 2010. *"Diseño e Implementación de un almacén de datos para la red nacional de Genética Médica".* Ciudad de la Habana : s.n., 2010.

Mavilio, Alfredo Sifontes. 2009. *Diseño del Datamart del subsistema de Conducidos.* 2009.

Microsoft Corporation. 2012. [En línea] 2012. [Citado el: 03 de 12 de 2012.]

Navarro, Jorge Rubio y Salinas2, José Manuel. 2002. Almacén de Datos para el Análisis y Difusión de la. *Facultad de Turismo.* España : s.n., 2002.

ONE . 2011. Portal de la estadistica dominicana. *ONE pone en línea Almacén Central de Datos del Sistema Estadístico Nacional.* [En línea] 11 de febrero de 2011. [Citado el: 03 de 12 de 2012.] http://www.one.gob.do/index.php?module=articles&func=display&aid=1377.

Postgres. 2005. [En línea] 2005. [Citado el: 02 de febrero de 2013.] http://www.postgresql.org/about/news/309/.

PostgreSQL. 2005. PostgreSQL. [En línea] 07 de 04 de 2005. [Citado el: 02 de 05 de 2013.] http://www.postgresql.org/about/news/309/.

Revista de Arquitectura e Ingeniería. **González Hidalgo-Gato, Ing. Gisel y Rizo Rizo, Lic. Emma. 2010.** 2, Cuba : Redalyc, 2010, Vol. 4.

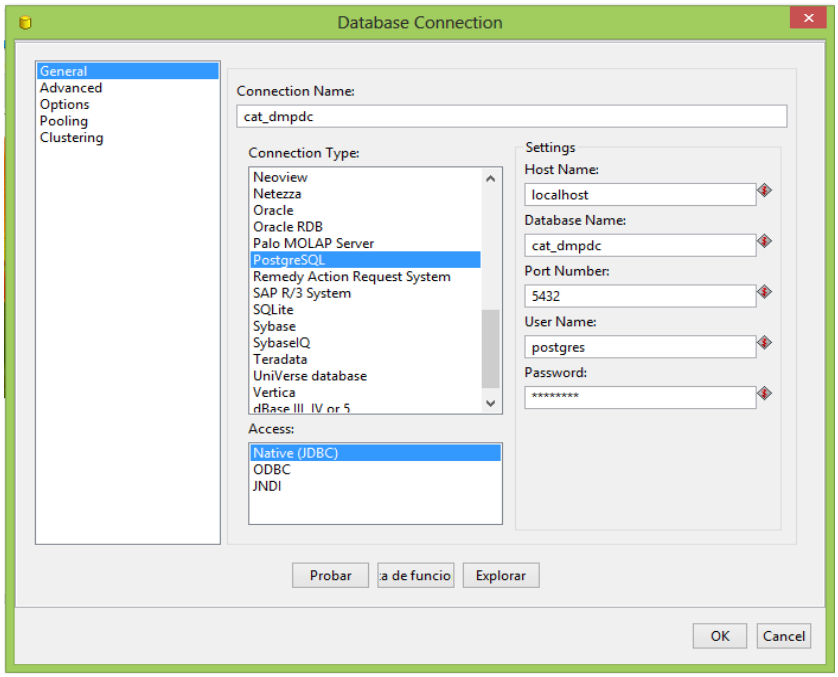
Rivadera, Gustavo R. 2010. *La metodología de Kimball para el diseño de almacenes de.* Buenos Aires : s.n., 2010.

summan. 2006-2008. Soluciones y servicios de manejo documental e infraestructura informática. [En línea] 2006-2008. [Citado el: 03 de 12 de 2012.] http://www.summan.com/pentaho/pentaho-bi-platform-server.

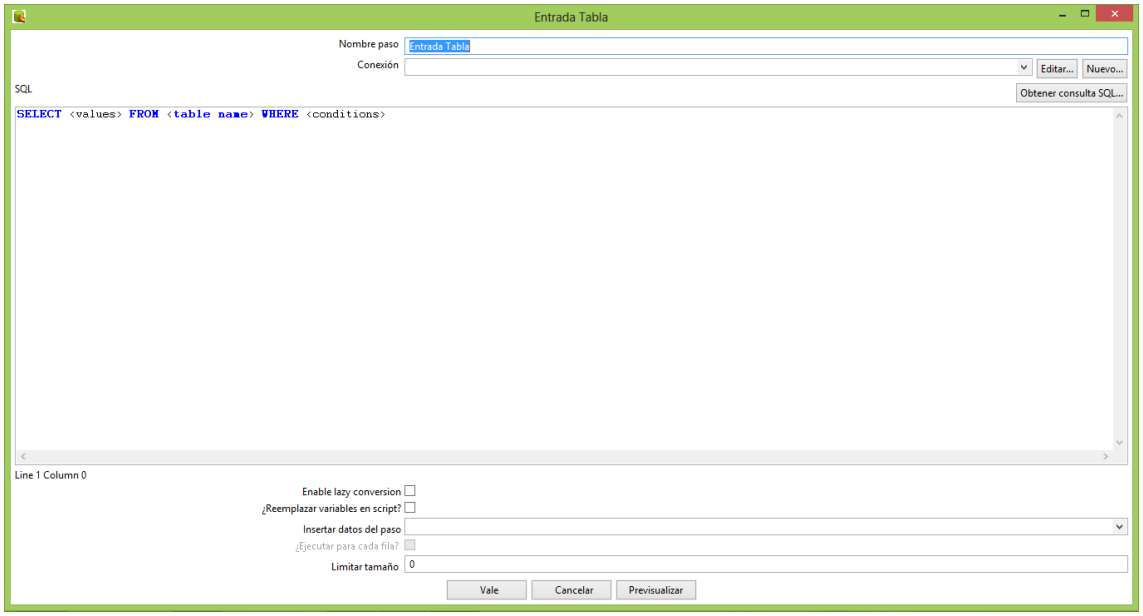
Targetware Informática S.A.C. . 2007-20012. Targetware . [En línea] 2007-20012. [Citado el: 07 de diciembre de 2012.] http://www.software.com.ar/visual-paradigm-para-uml.html.

Anexos:

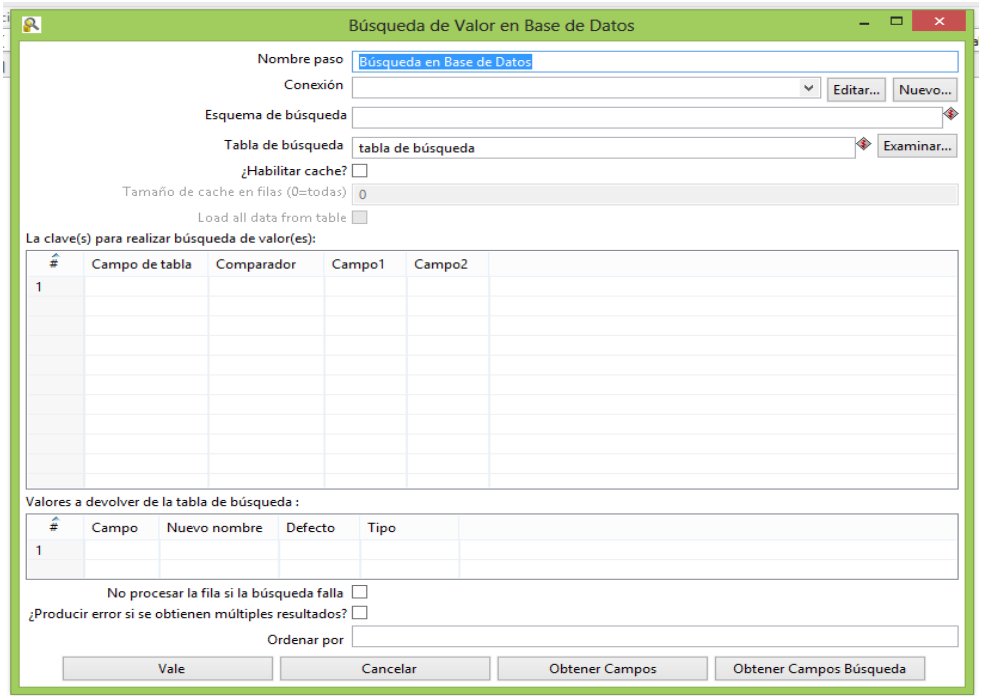
Anexo 1. Conexión con base de datos PostgreSQL en Kettle Pentaho Data Integration.



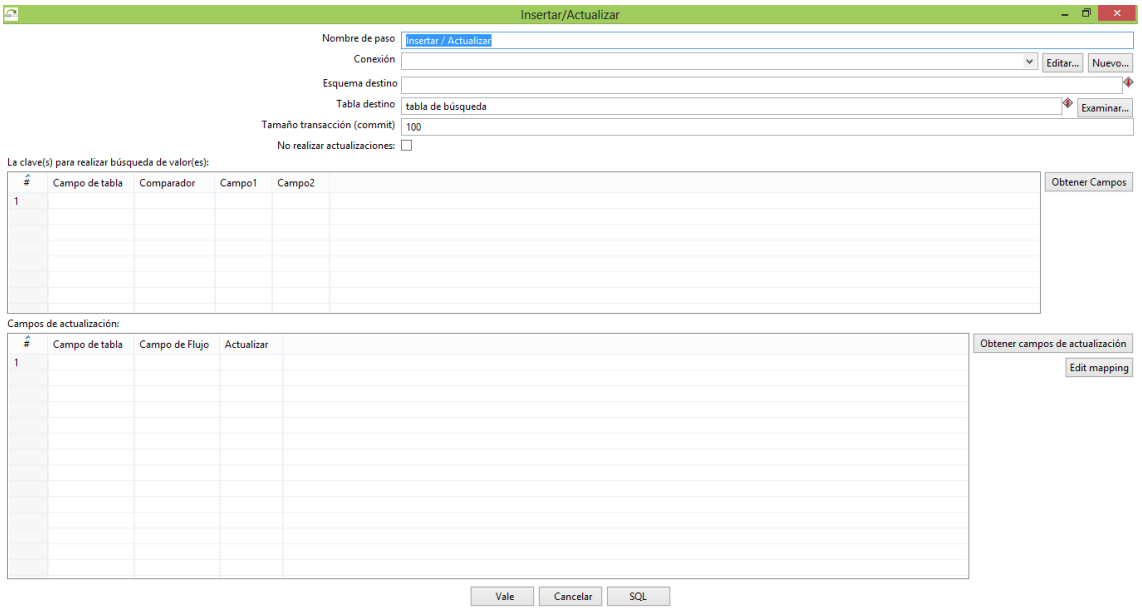
Anexo 2: Componente para la entrada de datos



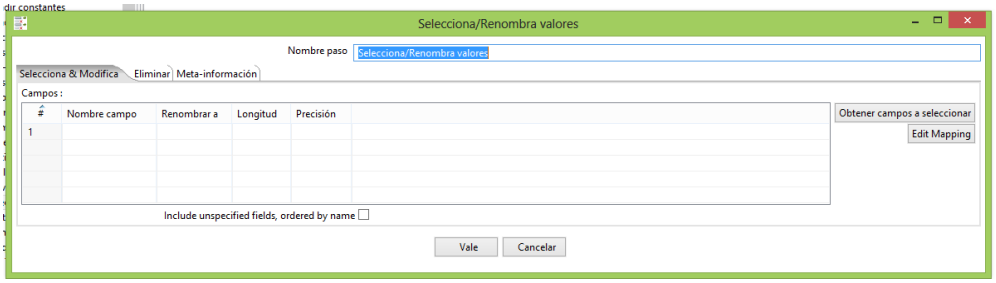
Anexo 3. Componente para realizar búsqueda en base de datos



Anexo 4. Componente para insertar o actualizar los valores



Anexo 5: Seleccionar y renombrar valores



Anexo 6: Transformaciones realizadas:

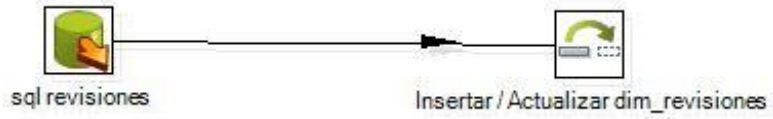


Ilustración 20 Transformación para la dimensión "dim_revisiones"

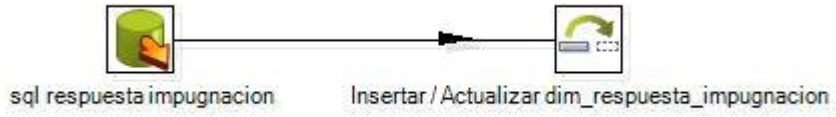


Ilustración 21 Transformación para la dimensión "dim_respuesta_impugnacion"

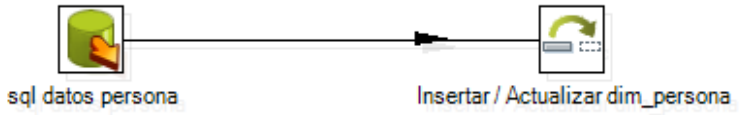


Ilustración 22 Transformación para la dimensión "dim_persona"

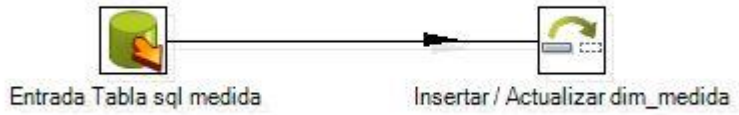


Ilustración 23 Transformación para la dimensión "dim_medida"

Ilustración 24 Transformación para la dimensión "dim_materia"

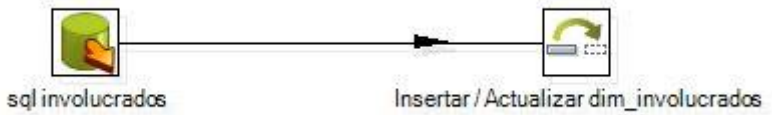


Ilustración 25 Transformación para la dimensión "dim_involucrados"

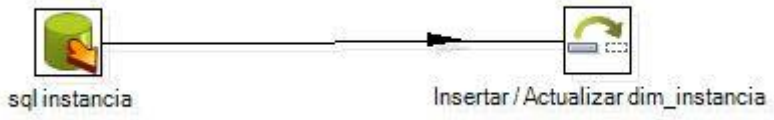


Ilustración 26 Transformación para la dimensión "dim_instancia"

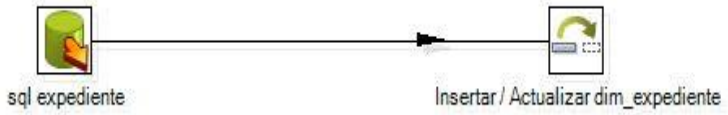


Ilustración 27 Transformación para la dimensión "dim_expediente"

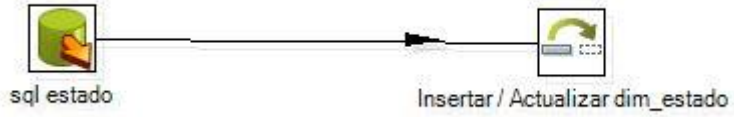


Ilustración 28 Transformación para la dimensión "dim_estado"

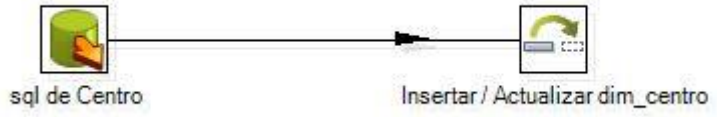


Ilustración 29 Transformación para la dimensión "dim_centro"

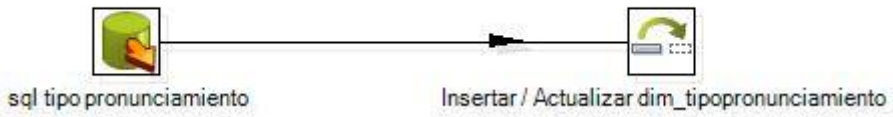


Ilustración 30 Transformación para la dimensión "dim_tipopronunciamiento"

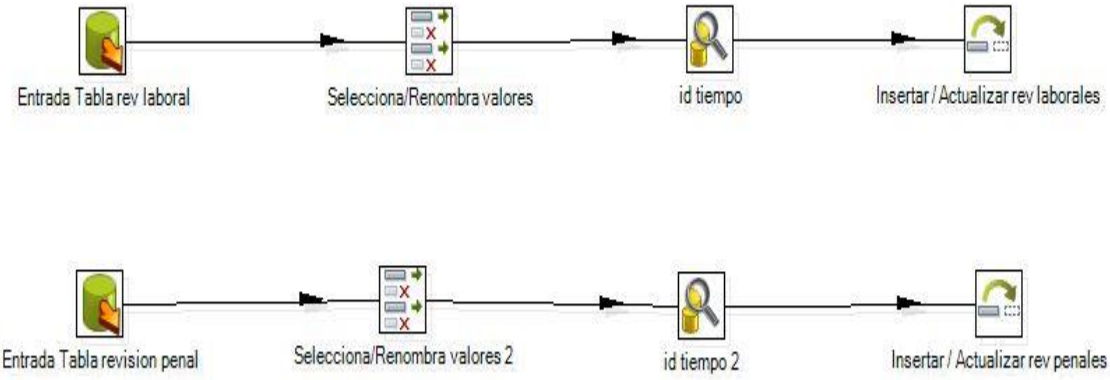


Ilustración 31 Transformación para el hecho "hecho_revisiones"

Ilustración 32 Transformación para el hecho "hecho_visitas"

Ilustración 33 Transformación para el hecho "hecho_quejas_reclamaciones"

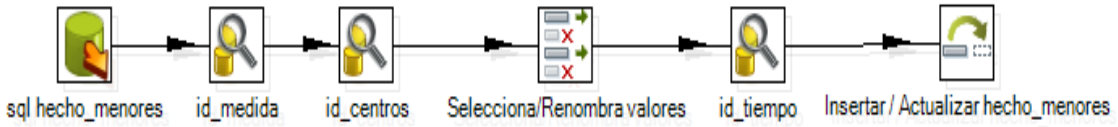


Ilustración 34 Transformación para el hecho "hecho_menores"

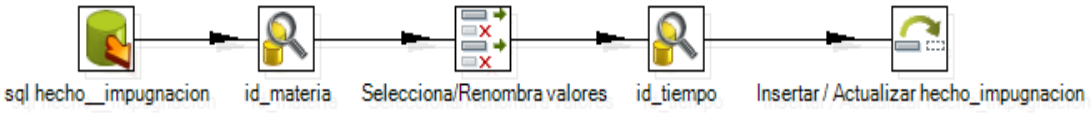


Ilustración 35 Transformación para el hecho "hecho_impugnacion"

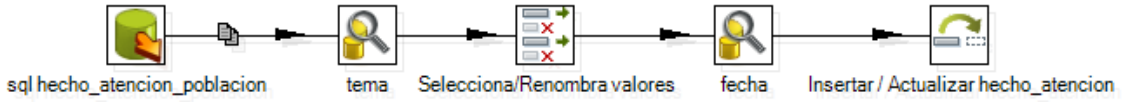


Ilustración 36 Transformación para el hecho "hecho_atencion_poblacion"

Ilustración 37 Transformación para la dimensión "dim_motivoacogida"

Anexo 7: Consultas.

Anexo 7.1 Consulta SQL para obtener los datos relacionados con la dimensión respuestaimpugnacion:

```
SELECT nresp.id_respuesta_impugnacion, nresp.descripcion from pdc.descritoimpugnacion
escimp INNER JOIN pdc.nrespuestaimpugnacion nresp
on escimp.id_respuesta_impugnacion = nresp.id_respuesta_impugnacion
```

Anexo 7.2: Consulta SQL para obtener los datos relacionados con la dimensión tipopronunciamiento:

```
SELECT pron.id_pronunciamiento, pron.descripcion from
pdc.drespuesta resp INNER JOIN pdc.ntipopronunciamiento pron
on resp.id_pronunciamiento = pron.id_pronunciamiento
```

Anexo 7.3: Consulta SQL para obtener los datos relacionados con la dimensión Revisiones:

```
SELECT sol.id_proceso, 'Revision Penal' as tipo_revision  from
pdc.dsolicitud sol INNER JOIN base.dproceso pro on sol.id_proceso = pro.id_proceso
GROUP BY sol.id_proceso
UNION
/*revisiones laborales*/
SELECT  dep.id_proceso, 'Revision Laboral' as tipo_revision FROM
base.dproceso dep INNER JOIN base.ntipoproceso tipo
on dep.id_tipo_proceso  = tipo.id_tipo_proceso
WHERE tipo.id_tipo_proceso = 4
GROUP BY dep.id_tipo_proceso,dep.id_proceso
```

Anexo 7.4: Consulta SQL para obtener los datos relacionados con la dimensión motivoacogida:

```
SELECT  mot.id_motivo_i,
(case when mot.descripcion IS  NULL then 'No definido' else mot.descripcion end) as
descripcion
 from  pdc.dinstitucionmenor_nmotivoingreso dimot
INNER JOIN pdc.nmotivoingreso mot on dimot.id_motivo_i = mot.id_motivo_i
```

Anexo 7.5: Consulta SQL para obtener los datos relacionados con la dimensión involucrados:

```
SELECT
(case when  dep.promueve_entidad=TRUE then 1 else 2 END ) as idinvolucrados,
```

```
(case when dep.promueve_entidad=TRUE then 'Promoventes' else 'Entidades Infractoras'
END ) as involucrados from pdc.descritoimpugnacion esc INNER JOIN
pdc.ddecisionadoptada des on esc.id_decision = des.id_decision INNER JOIN
pdc.ddepuracionescrito dep on des.id_decision = dep.id_decision
```

Anexo 7.6: Consulta SQL para obtener los datos relacionados con la dimensión expediente:

```
SELECT em.id_tipo_expte, te.descripcion from
pdc.dexpedientemenores em INNER JOIN pdc.ntipoexpediente te
on em.id_tipo_expte = te.id_tipo_expte
```

Anexo 7.7: Consulta SQL para obtener los datos relacionados con la dimensión centros:

```
SELECT dim.id_institucion,nti.descripcion as tipo
From pdc.dinstitucionmenor dim INNER JOIN
pdc.dinstitucionmenores dims on dim.id_institucion=dims.id_institucion inner JOIN
pdc.ntipoinstitucion nti on dims.id_tipo_institucion=nti.id_tipo_institucion
```

Anexo 7.8: Consulta SQL para obtener los datos relacionados con la dimensión estado:

```
SELECT DISTINCT ep.id_estado_paso, ne.descripcion from
base.nestado ne INNER JOIN
base.destadopaso ep on ne.id_estado = ep.id_estado INNER JOIN
base.dproceso dp on dp.id_estado_paso = ep.id_estado_paso INNER JOIN
base.npaso np on np.id_paso = ep.id_paso
```

Anexo 7.9: Consulta SQL para obtener los datos relacionados con la dimensión instancia:

```
SELECT concat(tip.descripcion, '_' , fis.denominacion ) as tipo_fiscalia,
fis.id_fiscalia FROM base.dfiscalia fis inner JOIN
base.ntipofiscalia tip on fis.id_tipo_fiscalia = tip.id_tipo_fiscalia
```

Anexo 7.10: Consulta SQL para obtener los datos relacionados con la dimensión respuestaimpugnacion:

```
SELECT nresp.id_respuesta_impugnacion, nresp.descripcion from
pdc.descritoimpugnacion escimp INNER JOIN pdc.nrespuestaimpugnacion nresp
on escimp.id_respuesta_impugnacion = nresp.id_respuesta_impugnacion
```

Anexo 7.11: Consulta SQL para obtener los datos relacionados con la dimensión motivo_acogida:

```
SELECT mot.id_motivo_i,
(case when mot.descripcion IS NULL then 'No definido' else mot.descripcion end) as
descripcion
 from pdc.dinstitucionmenor_nmotivoingreso dimot
INNER JOIN pdc.nmotivoingreso mot on dimot.id_motivo_i = mot.id_motivo_i
```

Anexo 7.12: Consulta SQL para obtener los datos relacionados con la dimensión persona:

```
SELECT per.id_persona,
(case when per.sexo IS NULL then 'No definido' else per.sexo end) as sexo,
(case when date_part('year', CURRENT_DATE)- date_part('year', per.anno_nacimiento) IS
NULL then 999 else
 date_part('year', CURRENT_DATE)- date_part('year', per.anno_nacimiento) end) as edad,
```

```
(case when per.raza IS  NULL then 'No definido' else per.raza end) as raza
 from base.dpersona per
```

Anexo 7.13: Consulta SQL para obtener los datos relacionados con la dimensión menor:

```
SELECT per.id_persona, per.sexo,  mot.descripcion,
date_part('year', CURRENT_DATE)- date_part('year', per.anno_nacimiento) as edad,
(case when per.raza IS  NULL then 'no definido' else per.raza end) as raza
 from
base.dpersona per INNER JOIN pdc.dinstitucionmenor inti on per.id_persona = inti.id_persona
INNER JOIN pdc.dinstitucionmenor_nmotivoingreso dimot on inti.id_institucion_menor =
dimot.id_institucion_menor
INNER JOIN pdc.nmotivoingreso mot on dimot.id_motivo_i = mot.id_motivo_i
WHERE
  (SELECT date_part('year', CURRENT_DATE)- date_part('year', per.anno_nacimiento) <= 18)
```

Anexo 7.14: Consulta SQL para obtener los datos relacionados con el hecho atención a la población:

```
SELECT atenc.id_proceso as idatencion,
count(perpro.id_persona) as cant_personas_atendidas,
fis.id_fiscalia as id_instancia,
concat(nt.descripcion,'_',sub.descripcion) as tema,
perpro.id_persona,
date(pro.fecha_creacion) as fecha
from
pdc.datencion atenc INNER JOIN base.dproceso pro on atenc.id_proceso = pro.id_proceso
INNER JOIN base.dpersonaproceso perpro on pro.id_proceso = perpro.id_proceso
INNER JOIN base.dfiscalia fis on  pro.id_fiscalia = fis.id_fiscalia
INNER JOIN base.dtipofiscalia_ntipodivision tipfi on fis.id_tipo_fiscalia = tipfi.id_tipo_fiscalia
INNER JOIN base.ntipofiscalia tipo on tipo.id_tipo_fiscalia = tipfi.id_tipo_fiscalia
INNER JOIN pdc.dplanteamiento plant on atenc.id_planteamiento = plant.id_planteamiento
INNER JOIN pdc.nsubtema sub  on plant.id_subtema = sub.id_subtema
INNER JOIN pdc.ntema_nsubtema ntst
on sub.id_subtema = ntst.id_subtema
INNER JOIN pdc.ntema nt on ntst.id_tema = nt.id_tema
GROUP BY fis.id_fiscalia, nt.descripcion,sub.descripcion, pro.fecha_creacion,
atenc.id_proceso,perpro.id_persona
```

Glosario de términos

Base de datos: También llamado banco de datos, es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso.

COUNT: Función definida en el lenguaje SQL para contar, también se puede encontrar definida en otros lenguajes.

Cubo: colección de dimensiones y medidas en un área temática particular.

MDX: es el acrónimo de MultiDimensional eXpressions, es un lenguaje de consulta para bases de datos multidimensionales sobre cubos OLAP.

Nivel de granularidad: La granularidad representa el nivel de detalle al que se desea almacenar la información sobre el negocio que se esté analizando.

Requerimiento: Es una necesidad documentada sobre el contenido, forma o funcionalidad de un producto o servicio.

Software: Es el conjunto de los programas de cómputo, procedimientos, regla, documentación y datos asociados que forman parte de las operaciones de un sistema de computación.

SQL: Es un lenguaje formal declarativo para manipular información en una base de datos.

SUM: Función del lenguaje SQL que devuelve el valor acumulado de una expresión.

UML: Lenguaje Unificado de Modelado (UML, por sus siglas en inglés, Unified Modeling Language) es el lenguaje de modelado de sistemas de software más conocido y utilizado en la actualidad.