UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 6

TITULO: “Sistema de Información de Gobierno”. Mercado de datos “Protección del Trabajo y Seguridad Social”.

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas.

AUTORES:

Leandro Gómez Herryman

Amauri Peña Cossío

TUTOR:

Ing. Vladimir Urquia Cordero

Ciudad de la Habana

Junio del 2011

Año 53 de la Revolución

Todos y cada uno de nosotros paga puntualmente su cuota de sacrificio consciente de recibir el premio en la satisfacción del deber cumplido, conscientes de avanzar con todos hacia el Hombre Nuevo que se vislumbra en el horizonte.

Ché

Declaración de autoría

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

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

_____ _____

Leandro Gómez Herryman

Autor

Amauri Peña Cossío

Autor

_____

Ing. Vladimir Urquia Cordero

Tutor

Datos de contacto

Tutor: Vladimir Urquia Cordero.

Categoría científica: Ingeniero en Ciencias Informáticas.

Correo: vurquia@uci.cu

Dedico este trabajo a quien fuera uno de mis grandes modelos a seguir, a la memoria de mi abuelo Francisco Roberto Gómez Fonseca.

Herryman

Dedico este trabajo a las dos personas más importantes que tengo en el mundo, aquellos que lo han dado todo para que mis planes futuros se cumplieran con éxito. Esas personas que han sido ejemplo a seguir durante mis años de estudio y han estado siempre ahí para cuando los necesito. A mis padres.

Amauri

Agradecimientos

A mis padres Rolando Gómez y Fe Margarita Herryman, por ser mis ejemplos y por apoyarme en todas las decisiones, que a lo largo de los años, me ha impuesto el destino. A mis hermanos Randy y Galber, por estar siempre a mi lado. A mis tíos Dulce María Y Eligio, por todo el apoyo y ayuda que me han dado durante estos años. A mí querido primo Adelkis, por toda la ayuda que me ha brindado. Al resto de mi familia, en especial a mi abuela Iranís. A los mejores amigos que he tenido, que me han apoyado en innumerables ocasiones: Alberto, Eugenio, Manresa, Luisito y en especial a Amauri. A la mamá de Amauri, por la ayuda brindad. En fin, a todos aquellos que de una forma u otra contribuyeron con el éxito del presente trabajo, les agradezco de todo corazón.

Herryman

A mis padres Rolando y Dania por estar siempre junto a mí, apoyándome en todas las decisiones que he tomado. A mi hermana Adriana, que aun siendo menor que yo siempre ha estado ahí recordándome que no importa lo mucho que te esfuerces, siempre puedes ser mejor. A mis abuelos Tita y Abu, por todo el apoyo y ayuda que me han dado durante estos años. A mi nené grande Isi, que aun estando en pruebas siempre me guardo parte de su tiempo y pude contar con su opinión y ayuda en el trabajo. A los amigos que he tenido durante estos cinco años: José, Ransel, Adrian, Alberto y un agradecimiento especial para Herryman, que aparte de ser un gran amigo, me ha ensenando que con empeño y dedicación siempre se obtienen los resultados esperados. En fin gracias a todas aquellas personas que de una forma u otra contribuyeron al éxito del presente trabajo.

Amauri

Agradecemos a los profesores Asnioby, Yanisbel y en especial a nuestro tutor Vladimir, que aun teniendo otras responsabilidades, pudieron dedicarnos parte de su tiempo.

Resumen

El presente Trabajo de Diploma se enmarca en el área de los Mercados de datos y la técnica de Procesamiento Analítico en Línea (OLAP, por sus siglas en inglés) para el análisis de información estadística. Comprende una revisión minuciosa y detallada de las metodologías, tendencias y mejores prácticas para el desarrollo de este tipo de soluciones, tomando como referencia el modelo de "Desarrollo de Soluciones de Almacenes de datos e Inteligencia de Negocio" formulado por el Centro de Tecnologías de Gestión de Datos (DATEC).

Como resultados se exponen las estructuras dimensionales para el modelo estadístico de Protección del Trabajo y Seguridad Social que comprende las dimensiones, jerarquías, tablas de hechos y medidas necesarias para soportar los análisis estadísticos. Además se definen los mecanismos de extracción, transformación y carga de los datos, así como la construcción de una capa de presentación, correspondientes al modelo. Otro aporte de la solución está referido a la seguridad, confiabilidad e integridad de los datos, definiendo niveles de seguridad, tanto a nivel de Base de Datos como a nivel de aplicación.

Por otra parte se presentan las estrategias de indexado y particionamiento que constituyen un referente para incorporar nuevos modelos. Su principal novedad consiste en su implementación utilizando herramientas libres.

Palabras claves:

- Oficina Nacional de Estadísticas (ONE).
- Almacén de datos.
- Mercado de datos.
- Sistema de Información de Gobierno (SIGOB).
- Base de Datos dimensional.
- Tabla de hechos.
- Tabla de dimensiones o dimensional.
- Medida.
- Expresión Multidimensional.
- Reporte.
- Transformación.
- Cubo OLAP o multidimensional.

Índice

Índice de ilustraciones y tablas

Ilustraciones

Tablas

Introducción

En las últimas décadas ha habido un creciente desarrollo de las Ciencias de la Información que le impone al mundo una nueva forma de concepción para enfrentarse a los problemas que diariamente se le presentan. Esta nueva forma de conceptualizar las soluciones a estos problemas se van fusionando indiscutiblemente al aumento de la explotación de las Tecnologías de la Información y las Comunicaciones (TIC) en la sociedad, mostrándose como un requisito indispensable para lograr un desarrollo sostenible.

El control de los datos estadísticos dentro de la infraestructura de un país constituye el eslabón principal para la toma de decisiones en los diferentes sectores socioeconómicos. Cuba posee una larga historia en materia de estadística. La entidad rectora de este tema en el país es la Oficina Nacional de Estadísticas (ONE), la cual mediante su Sistema Estadístico Nacional (SEN), organiza, dirige, controla y regula esta actividad, siendo la responsable de gestionar los principales indicadores de Protección del Trabajo y Seguridad Social.

La ONE tiene una estructura institucional distribuida territorialmente en las provincias y municipios del país, las cuales son las encargadas de interactuar directamente con los Centros Informantes (CI), siendo estos, el último eslabón de la cadena de la actividad estadística. Todas estas oficinas tienen atención administrativa y metodológica por la ONE. La información que se recoge en esta institución se difunde principalmente en papel y en CD-ROM, provocando afectaciones para su digitalización, pérdidas de información y aislamiento; complejizándose la situación si se consideran los 6833 CI existentes a todo lo largo del país.

Los CI por su parte han generado, con el pasar de los años, un histórico de datos disponibles en los más disímiles formatos. A medida que pasa el tiempo esa información se incrementa debido a la propia gestión estadística y aunado a esto, los avances tecnológicos reflejados en las redes y las telecomunicaciones, diversifican más esta situación.

Al realizar un análisis en los indicadores del modelo estadístico M5201 se detectaron las siguientes deficiencias en la presentación de los datos:

- Limitaciones para recuperar indicadores desde distintas perspectivas de análisis.
- Múltiples versiones de los datos, creando confusión a los analistas.
- Proceso de recuperación y elaboración de informes costosos en esfuerzo y tiempo.

Todo esto hace que el proceso de toma de decisiones sea un poco complejo y difícil de lograr.

El objeto social de la ONE conlleva a disponer de esta información de manera oportuna y con alta calidad para utilizarla como elemento de apoyo a la toma de decisiones a nivel nacional. La recuperación parcial o total de esta información, bajo las condiciones actuales, genera una gestión compleja y los resultados obtenidos no siempre se alcanzan en el tiempo y con la calidad requerida.

Partiendo de lo anteriormente expuesto se identifica el siguiente **problema de la investigación**: ¿Cómo contribuir a la toma de decisiones en el área "Protección del Trabajo y Seguridad Social" de la Oficina Nacional de Estadística?

Con el propósito de solucionar el mismo se traza como **objetivo general:** Desarrollar el Mercado de datos "Protección del Trabajo y Seguridad Social" del SIGOB (Sistema de Información de Gobierno) que contribuya a la toma de decisiones.

Para dar cumplimiento al **objetivo general**, los **objetivos específicos** propuestos son:

- Realizar el análisis y diseño del Mercado de datos del área "Protección del Trabajo y Seguridad Social".
- Implementar el Mercado de datos del área "Protección del Trabajo y Seguridad Social".
- Validar el Mercado de datos del área "Protección del Trabajo y Seguridad Social".

Por lo cual el **objeto de estudio** lo constituyen los Almacenes de datos y el **campo de acción** el Mercado de datos para al área "Protección del Trabajo y Seguridad Social" del SIGOB.

Para darle cumplimiento a los objetivos específicos trazados, se definieron las siguientes **tareas de investigación**:

- Caracterización de las metodologías, herramientas y tecnologías en el desarrollo de Almacenes de datos.
- Levantamiento de los requisitos.
- Descripción de los Casos de Uso del Mercado de datos.
- Definición de los hechos, las medidas y las dimensiones del Mercado de datos.
- Diseño del modelo de datos.
- Definición de la arquitectura de información del Mercado de datos.
- Diseño del subsistema de integración.

- Diseño del subsistema de visualización.
- Diseño de los casos de prueba.
- Implementación del modelo de datos.
- Implementación del subsistema de integración.
- Implementación del subsistema de visualización.
- Aplicación de las listas de chequeo.
- Aplicación de los casos de pruebas.

El presente documento está estructurado en cuatro capítulos, de los cuales el primero aborda la fundamentación teórica de los Mercados de datos, el segundo trata sobre el análisis y diseño del Mercado de datos “Protección del trabajo y seguridad social”, el tercero está orientado a la implementación del Mercado de datos, una vez realizado el análisis y el diseño del mismo, y el último capítulo expone lo concerniente a la validación del Mercado de datos implementado.

Capitulo 1. Fundamentación Teórica

Introducción

Desde un inicio, las Bases de Datos se convirtieron en una herramienta fundamental de control y manejo de operaciones comerciales. De ahí que en un corto período de tiempo las grandes empresas y negocios acumularan un cuantioso número de información que ya alcanzaba una dimensión considerablemente voluminosa. Con la acumulación de esta información se presentó la problemática de cómo darle un fin útil, debido a que en ella estaba almacenada la mayor parte de las operaciones comerciales de las mismas.

La solución sería unificar las diferentes fuentes de información de las cuales disponían, en un único lugar, al que sólo se le incorporaría información relevante, sobre la base de una estructura organizada, integrada, lógica, dinámica y de fácil explotación. La respuesta a esto fueron los Almacenes de datos o Data Warehouse, como se conocen mundialmente. (1), (2)

1.1 Almacén de datos

Existen diversas tendencias y formas de conceptualizar el término de Almacenes de datos, que aunque difieren en algunos aspectos, giran sobre el mismo eje central.

Al referirse a este particular; Claudia Imhoff, Nicholas Galemmo y Jonathan G. Geiger prestan especial atención al concepto desarrollado por Inmon en los años 90, en el que señala que los Almacenes de datos **“son un conjunto de datos orientados a un tema, integrados, de tiempo variante y no volátiles usados en la estrategia de toma de decisiones administrativas”**. Aseveran además que **“los Almacenes de Datos se han venido reconociendo cada vez más como una herramienta efectiva de las organizaciones para transformar los datos en información útil y estratégica para la toma de decisiones”**. (3)

Sin embargo Ralph Kimball, reconocido autor en el tema de Almacenes de datos, define un Almacén de datos como: **"una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis"**. Además determinó que un Almacén de datos no era más que: **"la unión de todos los Mercados de datos (del inglés Data Marts) de una entidad".** (4)

Como se puede apreciar, existe diversidad de puntos de vistas y percepciones sobre el tema. Aunque se expresan de forma diferente, queda claro que los Almacenes de datos son estructuras que se definen en función de temas específicos, donde la información histórica debe estar integrada, robusta

ante los cambios que puedan afectar a la organización y que su objetivo principal, que define su razón de ser, es servir de ayuda a la toma de decisiones empresariales.

1.2 Mercado de datos

Al hacer un análisis de la bibliografía existente sobre la tecnología de data warehousing, se puede constatar que existen autores que utilizan los conceptos de "**Almacén de datos**" y "**Mercados de datos**" indistintamente refiriéndose al mismo tema, aunque realmente no definen un concepto idéntico.

Los Mercados de datos son un subconjunto de datos de un Almacén de datos, donde se almacena la mayoría de las actividades de análisis que en el entorno de Inteligencia de Negocio se llevará a cabo. (3)

Un Mercado de datos es una alternativa de solución al igual que los Almacenes de datos a los problemas antes planteados, porque el diseño y construcción son similares, además de poseer una secuencia común. La diferencia entre estas dos estructuras se basa principalmente en que los Mercados de datos están enfocados en un área de negocio específica, mientras que un Almacén de datos entrega información a nivel corporativo. (5)

Un concepto más amplio sería: "Un conjunto flexible de datos, idealmente basado en el dato más atómico posible (granular) para ser extraído de las fuentes operacionales y presentado en un modelo simétrico (dimensional) que es más resistente cuando se enfrentan con las más inesperadas consultas de los usuarios (…) Podemos decir que los Mercados de datos están conectados con la arquitectura de los Almacenes de datos en su forma más simple y que representan los datos de un sólo proceso del negocio a la vez". (1)

1.3 Almacén de datos o Mercado de datos

En 1998 Bill Inmon declaró "El elemento más importante que enfrentan los directores de tecnologías de la información en este año es si construir primero los Data Warehouse". (6) Esta afirmación todavía está vigente en la actualidad.

Existen un conjunto de interrogantes que resultan imprescindibles evaluar antes de decidir realizar la construcción de un Almacén de datos: (6)

- ¿La aproximación se realizará de arriba hacia abajo (top-down) o de abajo hacia arriba (bottom-up)?
- ¿Empresarial o departamental?
- ¿Cuál primero el Almacén de datos o el Mercado de datos?

- ¿Construir un piloto o directamente el Almacén de datos completo?
- ¿Mercados de datos dependientes o independientes?

Las respuestas de estas preguntas conllevan a una planificación profunda de lo que realmente se va a desarrollar. Cada respuesta es crucial acerca de la decisión correcta sobre si utilizar un Almacén de datos Corporativo o utilizar un Mercado de datos Departamental.

En su libro "*Data Warehouse Fundamentals*", Poulraj Ponniah determina un conjunto de elementos que dan claridad sobre las diferencias existentes entre ambos conceptos. Estas diferencias se muestran a continuación en la **Tabla 1**: (6)

Almacén de datos	Mercado de datos
Corporativo o red empresarial.	Departamental.
Es la unión de todos los Mercados de datos.	Un simple proceso del negocio.
Los datos son recibidos desde el área de Procesamiento.	Unión en forma de estrella (hechos y dimensiones).
Consultas sobre la presentación de recursos.	Tecnología óptima para el acceso a los datos y el análisis.
Estructura para vista corporativa de los datos.	Estructura para adaptarse a la vista de los datos departamentales.

Tabla 1: Comparación entre Almacén de datos y Mercado de datos.

Tomando como base lo expresado por Ponniah, los Mercados de datos, como estructura, son un Almacén de datos pero reducido a un departamento específico sirviendo como fuente de análisis del tema que concierne a dicho departamento. La unión de todos los Mercados de datos de la organización es lo que conformaría la vista global de los datos, es decir, el Almacén de datos Corporativo.

Por lo tanto, la solución que se presenta es un Mercado de datos ("Protección del Trabajo y Seguridad Social" para SIGOB y no un Almacén de datos.

1.4 Características de los Mercados de datos

Inmon define un Almacén de datos en términos de las características del repositorio de datos. (2) Los Mercados de datos poseen las mismas características que los Almacenes de datos, las cuales son:

- **Integrado:** los datos almacenados en el Almacén de datos deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.
- **Temático:** sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. Por ejemplo, todos los datos sobre clientes pueden ser consolidados en una única tabla del Almacén de datos. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.
- **Histórico:** el tiempo es parte implícita de la información contenida en un Almacén de datos. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el Almacén de datos sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el Almacén de datos se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.
- **No volátil:** el almacén de información de un Almacén de datos existe para ser leído, pero no modificado. La información es por tanto permanente, significa que la actualización del Almacén de datos, es la incorporación de los últimos valores que tomaron las distintas variables contenidas en él, sin ningún tipo de acción sobre lo que ya existía.

Otra de las características de los Mercados de datos es que contienen metadatos (datos sobre los datos), los cuales permiten saber la procedencia de la información, su periodicidad de actualización, fiabilidad, forma de cálculo, etc. Además, permiten simplificar y automatizar la obtención de la información desde los sistemas operacionales a los sistemas informacionales. (4)

1.5 Subsistemas del Mercado de datos

Los Mercados de datos están integrados por tres subsistemas esenciales: el subsistema de almacenamiento, donde se gestionan los datos; el subsistema de integración, donde se realiza el proceso de Extracción, Transformación y Carga de los datos (ETL); y el subsistema de visualización, encargado de presentar los datos a los usuarios finales, a través de reportes e informes.

1.5.1 Subsistema de almacenamiento

1.5.1.1 Modelo Entidad Relación y Modelo Dimensional

Al realizar el modelo o diagrama de un Mercado de datos, surge la interrogante de si realizar un modelo Relacional o un modelo Dimensional.

Modelo Entidad-Relación

Un diagrama o modelo entidad-relación (a veces denominado por su siglas, *E-R* "Entity relationship", o, "DER", Diagrama de Entidad Relación) es un lenguaje para el modelado de datos de un sistema de información. Estos modelos expresan las entidades más relevantes para el sistema, sus inter-relaciones y propiedades. Trabajan dividiendo los datos en muchas entidades discretas donde cada una se convierte en una tabla física en la Base de Datos operacional. (7)

Los modelos Entidad–Relación no son recomendables para el diseño de los Almacenes de datos, debido a que no garantizan la recuperación óptima del gran cúmulo de información que se almacena. Además estos diagramas tienden a resultar en un diseño normalizado mientras que en un Almacén de datos este aspecto no es un requisito a tener muy en cuenta. (8)

Modelo Dimensional

A diferencia de los clásicos sistemas de Bases de Datos que presentan sus estructuras diseñadas mediante el modelo Entidad-Relación, los Almacenes de datos se diseñan mediante un modelo dimensional. Poseen la misma información que el DER, pero la organiza de forma diferente para garantizar la velocidad y eficiencia en la recuperación de la misma. Una de sus características principales es que no necesita una predefinición de los reportes, debido a que se diseñan de forma tal que cubra el universo de variantes que los usuarios necesiten consultar en la información almacenada. En la **Ilustración 1** se muestra la estructura espacial que posee este tipo de diseño.

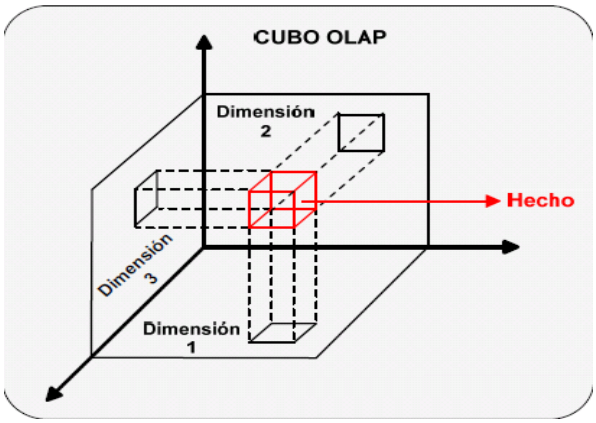


Ilustración 1: Estructura de un Cubo OLAP.

Para la materialización física de este tipo de modelo se utiliza comúnmente la propuesta realizada por Ralph Kimball llamada “esquema estrella”, que consiste en una tabla central denominada “tabla de hechos” y un conjunto de pequeñas tablas, llamadas “dimensiones”, que se relacionan a esta tabla central. Se le denomina estrella por su similitud con una estrella natural, debido a que consta de una tabla de hechos central y de varias tablas de dimensiones relacionadas a esta, a través de sus respectivas claves o llaves primarias, ver **Ilustración 2**.

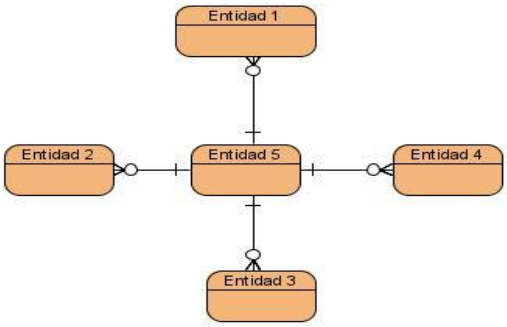


Ilustración 2: Representación del Esquema Estrella.

Existen otras estructuras que surgen producto de modificaciones realizadas al esquema estrella. En este sentido, se tiene el Copo de Nieve (*Snowflake*, en inglés)*;* la citada estructura tiene como primordial objetivo su uso para el ahorro de espacio de almacenamiento. Se plantea que una dimensión se encuentra “snowflaked” cuando los atributos de baja calidad se llevan a tablas separadas. La utilización de este tipo de estructura posee algunas deficiencias debido a que las presentaciones se tornan más complejas y afecta el rendimiento de la recuperación de las consultas. A este conjunto de estructuras anteriormente descritas se les puede añadir la Constelación de Hechos, donde su principal característica es que múltiples tablas poseen las mismas dimensiones. De esta forma se pueden utilizar diversas medidas, separadas en diferentes tablas de hechos, definidas por las mismas dimensiones.

El modelo dimensional divide el mundo de los datos en dos grandes conjuntos: las medidas y las descripciones del entorno de estas medidas. Las medidas, que generalmente son numéricas, se almacenan en las tablas de hechos y las descripciones de los entornos, que son textuales, se almacenan en las tablas de dimensiones.

Tablas de Hechos

La tabla de hechos es la tabla primaria en el modelo dimensional, donde el rendimiento de las mediciones numéricas del negocio es almacenado. (9) Generalmente cada tabla de hechos define un

Mercado de datos determinado debido a que en ellas se almacena la información concerniente al tema en cuestión, por ejemplo: ventas, clientes, vendedores, etc.

La principal condición que deben cumplir las tablas de hechos es que el hecho debe almacenarse de tal forma que su valor sea numérico y a su vez sea aditivo, para así poder realizar cálculos sobre él, ya sea por ciento, sumas, igualdades, etc.

Tablas de Dimensiones

Las tablas de dimensiones son las compañeras integrales de las tablas de hechos y son las que contienen la descripción textual del negocio. En el modelo dimensional, las tablas de dimensiones poseen varios atributos que en su conjunto definen una fila en la tabla de dimensión.

Los atributos de las dimensiones sirven como fuente primaria de las restricciones de las consultas, agrupaciones y las etiquetas de los reportes. Ellos desempeñan un rol de vital importancia dentro del Almacén de datos debido a que son las llaves que hacen al almacén usable y entendible. A su vez, son las llaves de entrada a los hechos o medidas almacenadas.

La calidad de todo Almacén de datos se mide por la definición de los atributos de las dimensiones. Su poder es directamente proporcional a la calidad y profundidad de estos atributos. (1)

1.5.1.2 Sistemas de almacenamiento

La forma de almacenamiento es crítica para garantizar el rendimiento de las consultas, las zonas de ubicación de las agregaciones y el procesamiento en general. Existen dos sistemas de almacenamiento diferentes, uno es el OLAP y el otro es el OLTP.

Sistemas OLAP

Los sistemas OLAP (On-Line Analytical Processing) son Bases de Datos orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos. Este sistema es típico de los Mercados de datos.

Sistemas OLTP

Los sistemas OLTP (On-Line Transactional Processing) son Bases de Datos orientadas al procesamiento de transacciones. Una transacción genera un proceso atómico (que debe ser validado con un *commit*, o invalidado con un *rollback*), y que puede involucrar operaciones de inserción,

modificación y borrado de datos. El proceso transaccional es típico de las Bases de Datos operacionales.

OLAP versus OLTP

La elección de uno u otro sistema depende de los requerimientos y las necesidades del sistema. En el **Anexo 1** se muestra una comparación de ambos sistemas de almacenamiento, según el acceso, estructuración, historial y organización de los datos.

Partiendo de la comparación de ambos sistemas de almacenamiento y teniendo en cuenta los requerimientos y necesidades identificadas en el sistema y del cliente, se aplica el sistema de almacenamiento OLAP.

1.5.1.3 Arquitecturas de almacenamiento

Existen tres arquitecturas de almacenamiento de la información para el proceso analítico en línea. La diferencia entre una u otra está dada por el modo en que son almacenados los datos.

Arquitectura ROLAP

La arquitectura ROLAP (Relational On-line Analytical Processing), accede a los datos almacenados en un Almacén de datos para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las Bases de Datos relacionales. (4)

Arquitectura MOLAP

La arquitectura MOLAP (Multidimensional On-Line Analytical Processing) usa Bases de Datos multidimensionales para proporcionar el análisis. Su principal premisa es que el OLAP está mejor implantado almacenando los datos multidimensionalmente. Este sistema usa una Base de Datos propietaria multidimensional, en la que la información se almacena multidimensionalmente, para ser visualizada en varias dimensiones de análisis. (4)

Arquitectura HOLAP

Un desarrollo un poco más reciente ha sido la solución OLAP híbrida (HOLAP “Hybrid On-Line Analytical Processing”), la cual combina las arquitecturas ROLAP y MOLAP para brindar una solución con las mejores características de ambas: desempeño superior y gran escalabilidad. Un tipo de HOLAP mantiene los registros de detalle (los volúmenes más grandes) en la Base de Datos relacional, mientras que mantiene las agregaciones en un almacén MOLAP separado. (4)

ROLAP versus MOLAP

La selección de un modelo en específico depende de cuán importante sea el rendimiento de las consultas para los usuarios y de la tecnología disponible a utilizar. En el modelo ROLAP la respuesta a las consultas y el tiempo de procesamiento suelen ser más lentos que con los modos de almacenamiento MOLAP u HOLAP. No obstante, ROLAP permite a los usuarios ver los datos en tiempo real y ahorrar espacio de almacenamiento, al trabajar con conjuntos de datos grandes a los que no se suele consultar con frecuencia, por ejemplo, datos puramente históricos. (10)

Por otro lado, las implementaciones ROLAP son más escalables y son frecuentemente atractivas a los clientes, debido a que aprovechan las inversiones en tecnologías de Bases de Datos relacionales ya existentes en la organización. En las implementaciones MOLAP el acceso a la información almacenada se realiza de forma más rápida y efectiva, utilizándose un depósito donde el tiempo en la velocidad de respuesta es crítico. Normalmente se desempeñan mejor que la tecnología ROLAP, pero tienen problemas de escalabilidad. En el **Anexo 2** se muestra una comparación entre ambos modelos basándose en Almacenamiento de los Datos, Tecnologías Subyacentes, Funciones y Características. (6)

Para el desarrollo del Mercado de datos (Protección Trabajo y Seguridad Social) se utiliza el sistema ROLAP por las ventajas que representa en este tipo de entorno, donde el detalle y la consolidación de los distintos conceptos relacionados condicionan el funcionamiento del sistema.

1.5.1.4 Herramienta de modelado

En la actualidad existen varias herramientas que facilitan el modelado de los datos, entre las que se encuentran Visual Paradigm for UML, ER-Studio, Rational Rose, etc.

La herramienta que se utiliza en el modelado de la solución del Mercado de datos es **Visual Paradigm for UML v 6.4**, por ser una herramienta CASE (Computer Aided Software Engineering) profesional para el desarrollo de aplicaciones, que integra el diseño visual, la generación de código fuente, Bases de Datos y documentación, abarcando todo el ciclo de vida del software. Entre las facilidades de uso que ofrece, es necesario mencionar la ayuda para asegurar la consistencia en el nombrado, las definiciones de tablas y columnas y la generación de los objetos físicos mediante el lenguaje DDL (Data Definition Language). Además, provee mecanismos para estimar las consecuencias de los cambios y su impacto en los diagramas de análisis, así como la integración con múltiples herramientas de desarrollo.

1.5.1.5 Sistema Gestor de Base de Datos

Los Sistemas Gestores de Bases de Datos (SGBD) son un tipo de software muy específico, dedicado a servir de interfaz entre la Base de Datos, el usuario y las aplicaciones que la utilizan. El propósito general de los sistemas de gestión de Base de Datos es el de manejar de manera clara, sencilla y ordenada un conjunto de datos que posteriormente se convierten en información relevante. (11)

Siguiendo la política nacional de migración hacia la independencia tecnológica, DATEC utiliza como SGBD PostgreSQL. Esta decisión ha sido previamente colegiada y aceptada por parte del cliente debido a que dentro de sus políticas de migración, se encuentran las de llevar todas sus Bases de Datos hacia dicha plataforma. En este sentido la versión seleccionada es la 8.4 por ser lo suficientemente estable y segura. Entre las principales características que avalan la decisión de usar PostgreSQL como SGBD figuran las siguientes: (12)

- Se ejecuta en más de 30 plataformas diferentes.
- Excelente documentación.
- Es sumamente adaptable a las necesidades propias.
- Soporta casi toda la sintaxis SQL. Soporte para los tipos de datos de SQL92, SQL99, SQL2003 y parte del SQL2008.
- Soporta llaves foráneas, tipos de datos definidos por el usuario, secuencias, relaciones, uniones, vistas, reglas, triggers, y procedimientos almacenados en múltiples lenguajes.
- Usa una arquitectura proceso-por-usuario cliente/servidor. Hay un proceso maestro que se ramifica para proporcionar conexiones adicionales para cada cliente que intente conectar a PostgreSQL.
- Tiene soporte para varios lenguajes procedurales internos incluyendo un lenguaje nativo denominado PL/pgSQL. Este lenguaje es comparable al lenguaje procedural de Oracle, PL/SQL. Posee habilidad para usar Perl, Python, Ruby o TCL como lenguaje procedural embebido, además de C, C++ y, Java.
- Puntos de recuperación a un momento dado, tablespaces, replicación asincrónica, transacciones jerarquizadas (savepoints), copia de seguridad en línea.
- Un sofisticado analizador/optimizador de consultas.
- Soporta juegos de caracteres internacionales, codificación de caracteres multibyte.

1.5.2 Subsistema de Integración

1.5.2.1 Integración de datos

Los procesos de integración de datos están basados en la necesidad de aunar los datos pertenecientes a múltiples fuentes de datos con el fin de obtener de forma centralizada, una mirada única e integrada al problema en cuestión. En principio y como principal problema los datos son heterogéneos y se encuentran distribuidos y dispersos, en la mayoría de los casos no estandarizados. Existen de esta forma numerosas islas de información inconsistentes que imposibilitan una comprensión unificada en cuanto a los términos, cantidades, unidades de medida, etc. Las entidades generadoras de datos, provocan que la tarea de unificar estos datos sea sumamente compleja y costosa. Es por ello que al hablar de integración de datos aparece el término calidad de datos, como una rama o subproceso a tener en cuenta. La información proveniente de diferentes sistemas es entonces inconsistente y con baja calidad, la gran mayoría de las veces al intentar compatibilizarla con otros sistemas, incluso dentro de las mismas empresas o entidades, es común encontrar múltiples sistemas, tecnologías, canales de comunicación y transporte cuando de integración de datos se trata. Tomando en consideración las diferencias entre la calidad de datos, los procedimientos para el manejo de datos de los expertos, la diferencia de formatos, lenguajes y muchos otros problemas es preciso señalar que la plena y armoniosa integración de los recursos de información es prácticamente un problema imposible de solucionar. Es por esto que la integración de datos es necesaria.

1.5.2.2 Proceso de ETL

El proceso de Extracción, Transformación y Carga (ETL) de los datos es el encargado de impulsar el flujo de datos haciendo transformaciones intermedias y permitiendo una integración de datos exitosa. Es por esto que cada paso, desde su diseño de acuerdo a los requerimientos de cada negocio, hasta su eficiente puesta en marcha, precisa de la mayor atención y esfuerzo posibles.

El proceso ETL permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos, limpiarlos, y cargarlos en otra Base de Datos, protegiendo el linaje de los mismos. ETL es un proceso enfocado a la integración de datos, tanto por lote, como en tiempo real hacia Almacenes de datos, logrando alto grado de transformaciones para la migración y consolidación. Este proceso sincroniza datos desde diversas aplicaciones e involucra procesos de manipulación que van más allá de un simple movimiento desde el punto A hasta el punto B; donde este proceso no está separado de los sistemas operacionales, sino que está integrado con los demás procesos de la empresa.

El proceso de ETL se compone de tres subprocesos fundamentales que permiten la división y un entendimiento efectivo de dicha labor.

Extracción de los datos

Consiste en extraer los datos desde los sistemas de origen. Estas fuentes primarias pueden encontrarse sobre arquitecturas, sistemas y estructuras heterogéneas. La mayoría de los proyectos de almacenamiento de datos fusionan datos provenientes de diferentes sistemas de origen. Cada sistema separado puede usar una organización diferente de los datos o formatos distintos. Los formatos de las fuentes normalmente se encuentran en Bases de Datos relacionales o ficheros planos, pero pueden incluir Bases de Datos no relacionales u otras estructuras diferentes. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación. Una parte intrínseca del proceso de extracción es verificar los datos extraídos, de lo que resulta un chequeo que comprueba si los datos cumplen la pauta o estructura que se esperaba. De no ser así los datos son rechazados. Un requerimiento importante que se debe exigir en la tarea de extracción es que esta cause un impacto mínimo en el sistema origen. Si los datos a extraer son muchos, el sistema de origen podría colapsar, provocando que no pueda utilizarse con normalidad para su uso cotidiano. Por esta razón, en sistemas grandes las operaciones de extracción suelen programarse en horarios o días donde este impacto sea nulo o mínimo. Las herramientas utilizadas en la extracción deben ser adaptables, extensibles y capaces de filtrar los datos relevantes a extraer de las fuentes, permitiendo la compresión, descompresión, y encriptación de datos.

Limpieza y Transformación de los datos

Etapa central del proceso, donde los datos extraídos son convertidos de su estado original a un formato consistente con el repositorio destino, sin perder su exactitud o veracidad con respecto a las fuentes. Para esta etapa medular, los datos deben ser limpiados, pues los datos en el mundo real son inconsistentes; se muestran incompletos, donde faltan valores de los atributos, ciertos atributos de interés, o contienen solo agregados de datos; se presentan con ruido, conteniendo errores o valores fuera de límites e inconsistentes, manteniendo discrepancias en nombre o códigos; por ello se realiza un proceso de limpieza que elimina errores e inconsistencias en los datos y resuelve el problema de identidad de los objetos. Este proceso tiene como tareas fundamentales: llenar valores ausentes, identificar valores fuera de límite y eliminar el ruido en los datos, corregir las inconsistencias, e integrar los mismos.

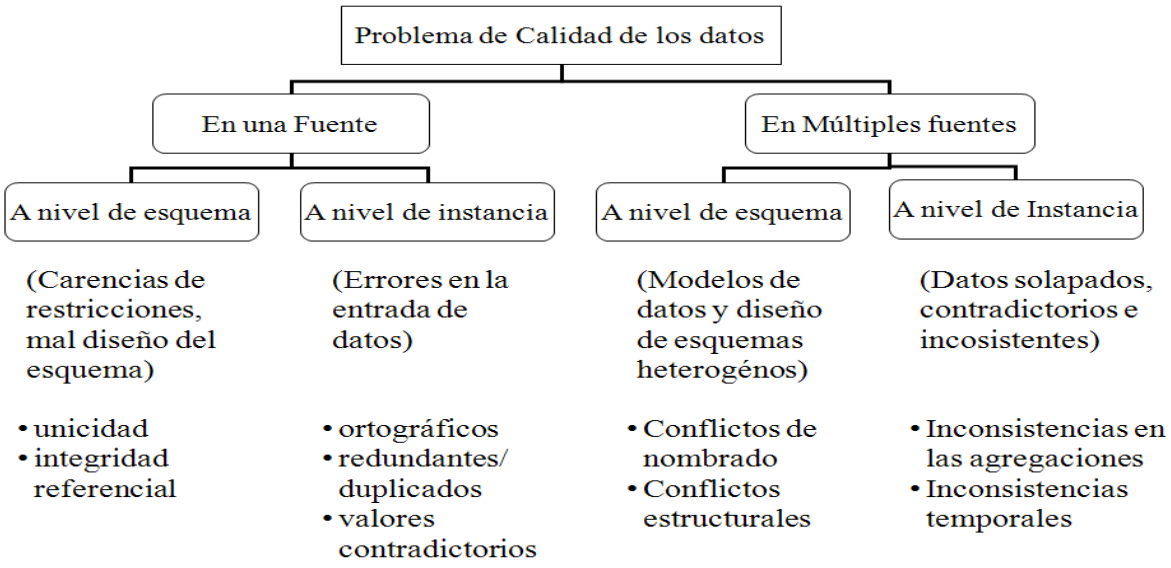


Ilustración 3: Problemas a resolver por la limpieza de datos.

Una vez que los datos se encuentren integrados se procede a realizar las transformaciones, que pueden ir desde simples conversiones de formato, hasta las más complejas operaciones de integración. Aunque lo idóneo es realizar los siguientes pasos:

- Análisis de datos.
- Definición del flujo de trabajo de las transformaciones y las reglas de correspondencia.
- Verificación.
- Transformación.
- Flujo inverso de datos limpios.

Carga de los datos

Es el momento donde se realiza la transferencia del conjunto resultado de las transformaciones a su destino, ya sean sistemas o ficheros con cierto formato. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes. En algunas Bases de Datos se sobrescribe la información antigua con nuevos datos. Los Almacenes de datos mantienen un historial de los registros de manera que se pueda hacer una auditoría de los mismos y disponer de un rastro de toda la historia de un valor a lo largo del tiempo. La fase de carga interactúa directamente con la Base de Datos de destino. Al realizar esta operación se aplicarán todas las restricciones y triggers (disparadores) que se hayan definido. Estas restricciones y disparadores contribuyen a que se garantice la calidad de los datos en el proceso ETL.

1.5.2.3 Arquitectura de la solución ETL

La arquitectura de la solución ETL está integrada por varios componentes.

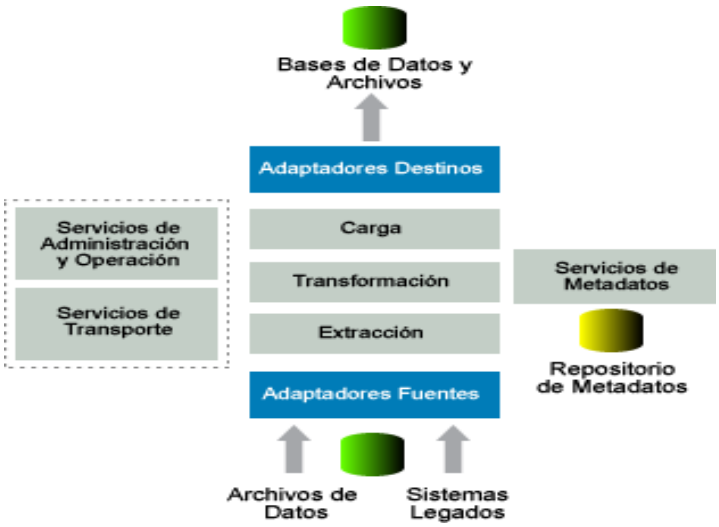


Ilustración 4: Arquitectura de la solución ETL.

Los componentes de la arquitectura mostrada son los siguientes:

- **Servicios de administración y operaciones:** estos servicios aseguran la utilización efectiva de los recursos en el ambiente de sincronización y una administración efectiva mediante la planificación y seguimiento de tareas, gestión de metadatos (datos sobre datos), recuperación de errores, etc.
- **Servicios de transportación:** procesos que garantizan el movimiento de la información cruda o transformada desde una fuente hasta un repositorio destino.
- **Servicios de metadatos:** Los metadatos son información descriptiva sobre los datos y otras estructuras, como objetos, reglas de negocio y los procesos que manipulan los datos. Los metadatos pueden ser agrupados en dos categorías:
    - **Metadatos mecánicos:** enfocados a los diseñadores, desarrolladores y administradores durante el desarrollo y mantenimiento del proceso. Este es el punto técnico que agrupa las herramientas, aplicaciones y sistemas, para que juntos constituyan la solución.
    - **Metadatos del negocio:** brindan una imagen clara del servicio del ambiente de trabajo a los usuarios finales.

1.5.2.4 Herramientas de Integración de datos

Las herramientas de ETL proporcionan consolidación de datos para la construcción de Bases de Datos permanentes utilizadas en el análisis o la generación de informes. Utilizan tres funciones principales combinadas en una herramienta que extrae datos de Bases de Datos fuentes y las coloca en Bases de Datos destino.

Pentaho Data Integration

Es una herramienta de código abierto adoptado por Pentaho BI. Proporciona la extracción de gran alcance, transformación y carga de los datos utilizando un enfoque innovador, orientado a los metadatos. Posee una interfaz intuitiva, gráfica de arrastre, una probada arquitectura escalable y basada en estándares. Puede crear complejas transformaciones y emplear un entorno gráfico sin tener que generar ningún código personalizado. Pentaho Data Integration es una solución de ETL con todas las funciones incluyendo:

- Rica colección de transformación con más de 150 objetos de asignación.
- Amplia fuente de datos de apoyo incluyendo aplicaciones empaquetadas, más de 30 plataformas de código abierto y propietario de Base de Datos, archivos planos, documentos de Excel, y otras.
- Apoyo al análisis de datos con la integración y gestión de datos.
- Rendimiento y escalabilidad.
- La integración con la suite de Pentaho BI para Enterprise Information Integration (EII), programación avanzada, y la integración de procesos.
- Unificado de ETL, modelado y visualización de entorno de desarrollo para el diseño de aplicaciones de BI.

DataCleaner

Es una aplicación Open Source para el perfilado, la validación y comparación de datos. Estas actividades ayudan a administrar y supervisar la calidad de los datos, garantizando la utilidad de la información a su situación de negocio. Es una de las aplicaciones más fáciles de usar para la calidad de los datos. Normalmente se utiliza antes, durante y después del proceso ETL.

- Antes, para profundizar en los orígenes de datos que serán usados en el trabajo. Por lo general se refieren a esto como mirar debajo de la punta del iceberg de los datos.
- Durante, en caso de existir desajustes inesperados durante el proceso de ETL.

- Después de asegurar la coherencia y la calidad en la fuente de datos que han poblado.
- DataCleaner puede acceder y analizar prácticamente cualquier Almacén de datos, incluyendo: base datos, hojas de cálculo Excel, archivos XML (Extensible Markup Language), etc.

1.5.3 Subsistema de visualización

1.5.3.1 Inteligencia de Negocios

La Inteligencia de Negocios o Bussiness Intelligence (BI), se conforma con el conjunto de una serie de herramientas y estrategias que orientadas a la administración y creación de conocimientos, es capaz de facilitar el proceso de toma de decisiones en determinada empresa o institución, mediante el análisis de los datos existentes. La Inteligencia de Negocios se refiere al uso de la tecnología para recolectar y usar efectivamente la información, a fin de mejorar la operación del negocio. Un sistema ideal de BI ofrece a los empleados, socios y altos ejecutivos, acceso a la información clave que necesitan para realizar sus tareas del día a día, y principalmente para poder tomar decisiones basadas en datos correctos y certeros. (13)

Se compone de una serie de herramientas y técnicas de ETL, o sea, información relacionada con la empresa o con determinada área de la misma. Estos datos, son extraídos, se depuran y preparan para ser cargados en un Almacén de datos.

1.5.3.2 Arquitectura de la solución BI

La solución de BI parte de los sistemas de origen de una organización (fuentes de datos). Las Fuentes de Datos son los diversos sistemas de almacenamiento existentes en la empresa (el sistema financiero, el sistema de almacenamiento, plantillas y archivos de texto, tablas, datos externos de Internet, paquetes empresariales como ERPs (Enterprise Resource Planning), CRMs (Customer Relationship Management)). Después de definir las Fuentes de Datos, la siguiente fase es realizar el proceso de ETL.

La información resultante, ya unificada, depurada y consolidada, sirve como base para la construcción del Mercado de datos “Protección del Trabajo y Seguridad Social”, el cual se caracteriza por poseer la estructura óptima para el análisis de los datos del área “Protección del Trabajo y Seguridad Social” de la ONE, mediante una Base de Datos analítica.

Los datos albergados en el Mercado de datos se explotan utilizando herramientas comerciales de análisis, reporting, alertas, etc. En estas herramientas se basa también la construcción de productos BI

más completos, como los sistemas de soporte a la decisión (DSS), los sistemas de información ejecutiva (EIS) y los cuadros de mando (CMI) o Balanced Scorecard (BSC). (14)

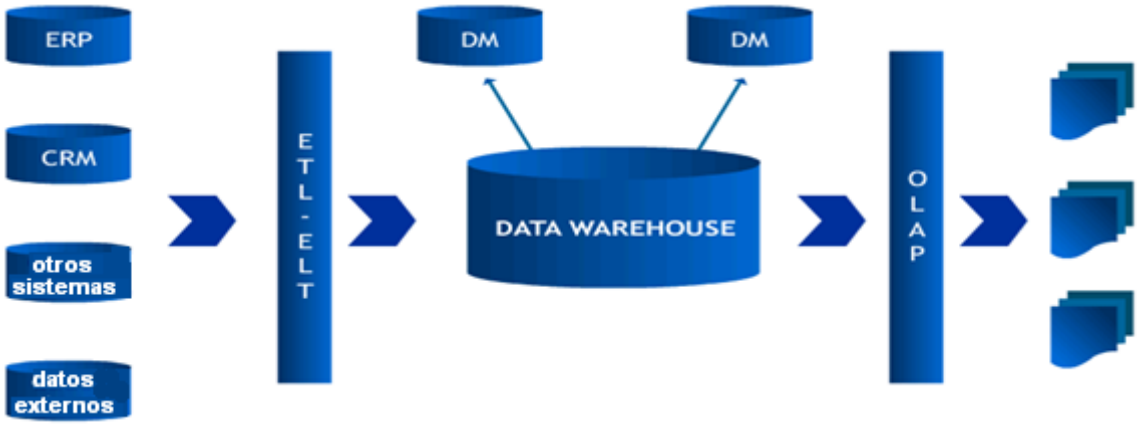


Ilustración 5: Arquitectura de la solución BI.

1.5.3.3 Herramientas de visualización

Para realizar la visualización del los datos se utilizan una serie de herramientas que facilitan este proceso.

Pentaho BI Server

La plataforma Pentaho BI Server provee el soporte y la infraestructura necesarios para crear soluciones de inteligencia empresarial a problemas de negocios. El marco proporciona los servicios básicos, incluidos autenticación, registro, auditoría, servicios web y motor de reglas. La plataforma también incluye un motor de solución que integra reportes, análisis, tableros de comandos y componentes de minería de datos. El diseño modular y arquitectura basada en plugin permite a todos o parte de la plataforma estar inmersa en aplicaciones de terceros por los usuarios finales, así como fabricantes de equipos originales.

La aplicación Pentaho BI Server funciona como un sistema basado en administración web de informes, el servidor de integración de aplicaciones y un motor de flujo de trabajo ligero (secuencias de acción.) Está diseñado para integrarse fácilmente en cualquier proceso de negocio.

Algunas de sus ventajas son:

- Integración con procesos de negocio.
- Administra y programa reportes.
- Administra seguridad de usuarios.

Workbench

Mondrian Schema Workbench es un entorno visual para el desarrollo y prueba de cubos OLAP Mondrian. Si bien la definición del XML para esquemas Mondrian no es extremadamente compleja, en la práctica resulta engorroso recordar cada uno de los elementos junto a sus atributos y sub-elementos. Con esta aplicación, se puede configurar una conexión JDBC como el modelo físico, para luego elaborar el esquema lógico de manera simple y efectiva. Permite crear y probar los cubos OLAP visualmente para que luego el motor de Mondrian procese las Expresiones Multidimensionales (MDX) con los esquemas creados. Los esquemas son modelos metadatos XML que se crean en una estructura específica utilizada por el motor de Mondrian.

Ofrece las siguientes funcionalidades:

- Editor de esquemas integrados con un origen de datos subyacente para su validación.
- Prueba de consultas MDX contra el esquema y la Base de Datos.
- Examinar la estructura subyacente de Bases de Datos.

Apache Tomcat

Es una implementación de software de código abierto de Java Servlet y tecnologías JavaServer Pages. Es desarrollado en un entorno abierto y participativo y publicado bajo la licencia Apache versión 2. Es la intención de ser una colaboración de los mejores desarrolladores de su clase de todo el mundo. Funciona en cualquier sistema operativo que disponga de una máquina virtual de java. Tomcat puede utilizarse como un contenedor solitario o como plugin para un servidor web existente.

La versión 6.0.x posee nuevas características, entre las que se encuentran:

- Conformidad de mantenimiento.
- Requiere un mínimo de JDK 1.5.
- web.xml ya no es necesario.
- Anotaciones de apoyo.
- Alternativas a algunas entradas XML en web.xml.
- Nuevo lenguaje de expresión unificada.
- Tiene su propia especificación.
- Nueva manipulación TLD.
- Inyección de recursos.
- API basado en j.u.concurrent.Executor.

- Nuevo conector Java NIO.
- Permite múltiples url-pattern en servlet-mapping.

1.6 Metodología de desarrollo

El vocablo metodología, en la ciencia que estudia los métodos del conocimiento, se refiere a los métodos o procedimientos de investigación que se siguen para alcanzar una gama de objetivos en una ciencia. Además puede categorizarse como el conjunto de métodos que se rigen en una investigación científica o en una exposición doctrinal. (7)

Existen dos criterios bien identificados y que han marcado claramente su tendencia sirviéndole de guía a la comunidad mundial en cuanto a este tema. Estas tendencias son la conocida como Metodología Kimball en honor a su creador Ralph Kimball e igualmente mencionada Metodología de Inmon dada a su creador William H. Inmon. Bill Inmon y Ralph Kimball son dos de las personalidades referentes y más influyentes en el área de data warehousing, y responsables de los dos enfoques a los que se hace referencia. Inmon es el creador del término Data Warehouse así como del CIF (*Corporate Information Factory*), conjuntamente con Claudia Imhoff. Por su parte, Ralph Kimball es un gurú del diseño de Almacenes de datos y creador del enfoque MD (*Multidimensional Architecture*).

La principal diferencia que existe entre ambas tendencias está basada en la forma de enfrentar el problema. En el **Anexo 3** se muestran las principales diferencias entre las dos tendencias. (15)

No se puede decir que una de estas dos tendencias es mejor que la otra ya que los beneficios de cada metodología depende del ambiente de la empresa en que se vayan a implantar. Basados en estas propuestas se han desarrollado un conjunto de metodologías que no siguen obligatoriamente una en específico sino que realizan una selección de lo mejor de cada una. Como ejemplo se presenta el modelo de “Desarrollo de Soluciones de Almacenes de datos e Inteligencia de Negocio” (DW&BI) creada por DATEC, que utiliza como base la Metodología Kimball por los siguientes elementos:

- Utiliza el concepto de Dimensiones y Hechos, lo cual facilita el proceso de desarrollo y lo hace más eficaz.
- Propone ir construyendo el Almacén de datos a través de la construcción de los Mercados de datos departamentales, lo que constituye una buena estrategia y coincide con la división lógica de las empresas, entidades y organismos.
- Es una metodología madura y reconocida por el resto de la comunidad dedicada al tema. Tiene bien definidas las etapas, actividades, artefactos y roles.

- Es importante mencionar que el modelo creado por DATEC es una propuesta resistente y adaptable ante los cambios. Dentro del ciclo de vida presenta una serie de flujos de trabajo que serán mencionados a continuación:
    - **Estudio Preliminar o Planeación:** Se realiza el estudio de la entidad cliente, la planeación del proyecto, se definen los objetivos, el alcance preliminar, los costos estimados y otras actividades.
    - **Requerimientos:** Se realiza en dos direcciones: una mediante la identificación de las necesidades de información y reglas del negocio y la otra con un levantamiento detallado de las fuentes de datos a integrar. Después se procede a la definición de los requerimientos.
    - **Arquitectura y Diseño:** Se definen las estructuras de almacenamiento, se diseñan las reglas de extracción, transformación y carga y se define la arquitectura de información que regirá el desarrollo de la solución.
    - **Implementación:** Se diseña físicamente el Repositorio de Datos, se crean las estructuras de almacenamiento, el Área Temporal de Almacenamiento, se ejecutan las reglas de ETL y se configuran e implementan las herramientas de BI para la obtención de los elementos (reportes, gráficos, etc.) que se acordaron con el cliente.
    - **Prueba:** Se realizan las pruebas al sistema desde las Pruebas de Unidad hasta las de Aceptación con el cliente.
    - **Despliegue:** Se realiza un Despliegue Piloto en el cual se configuran los servidores, se instalan las herramientas y se carga una muestra de los datos para demostrar que el sistema funciona. Posterior a la aceptación del cliente se realiza la carga de los datos y la Capacitación y Transferencia Tecnológica.
    - **Soporte y Mantenimiento:** Tras la implantación de la solución se brindan los servicios de soporte en línea, vía telefónica, web u otras según el contrato firmado y las condiciones de soporte establecidas.
    - **Gestión y Administración del Proyecto:** A lo largo del ciclo de vida se realizan actividades de control, gestión y chequeo del desarrollo, los gastos, las utilidades, los recursos y demás actividades por parte del Grupo de Dirección del Proyecto.

Como se ha descrito, en cada flujo intervienen grupos específicos, cada uno con actividades y responsabilidades concretas. A continuación se mencionan los mismos:

- Grupo de Análisis.
- Grupo de Almacenes de datos.
- Grupo de ETL.
- Grupo de BI.
- Grupo de Dirección.

En la siguiente ilustración se puede apreciar la representación de los grupos en cada flujo de trabajo.

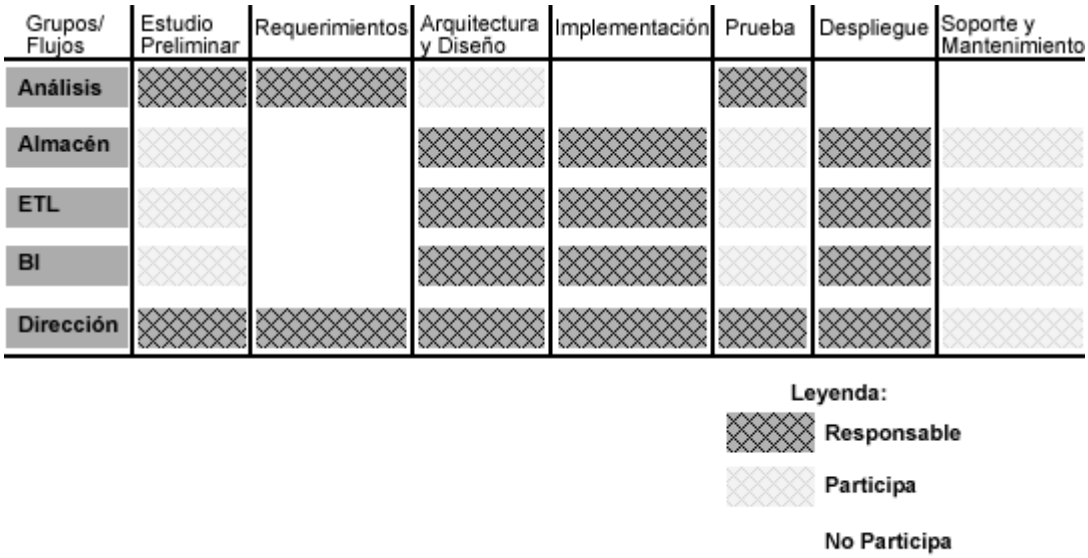


Ilustración 6: Grupos y Flujos de trabajos.

Conclusiones del capítulo

A partir del estudio del estado del arte realizado se concluye que:

- Los Mercados de datos son una solución viable para el almacenamiento de los datos pertenecientes al modelo M5201 “Indicadores Seleccionados de Protección del Trabajo y Seguridad Social” del Sistema de Información Estadística Nacional (SIEN).
- Se utiliza la topología Constelación de Hechos, la cual permite el manejo eficiente y organizado de la información para la generación de cubos OLAP y reportes multidimensionales.

- El modelo de “Desarrollo de Soluciones de Almacenes de datos e Inteligencia de Negocios” ha mostrado ser robusto y viable, potenciando el desarrollo eficiente y organizado en cada iteración del ciclo de desarrollo del Mercado de datos.
- Las herramientas seleccionadas viabilizan el desarrollo del Mercado de datos y hacen posible dar solución a todas las necesidades de los clientes. Además cumplen con la política de migración a software libre, siendo esto un aspecto indispensable para la situación económica y el momento histórico actual por el que cursa el país.

Capítulo 2. Análisis y Diseño del Mercado de datos

Introducción

El desarrollo de Mercados de datos no es una tarea sencilla debido a que está compuesto por tres componentes que interactúan entre sí como un sistema. Para lograr un diseño robusto y adaptable a las necesidades reales de los usuarios finales es necesario realizar un análisis exhaustivo del Mercado de datos, teniendo en cuenta elementos como: levantamiento de requisitos, reglas del negocio y los casos de uso del sistema.

2.1 Análisis del Mercado de datos

El análisis es la fase determinante para entender los requisitos de la organización, mediante el cual se definen las estructuras de almacenamiento y las reglas de transformación, así como la arquitectura de información que regirá el desarrollo de la solución.

2.1.1 Definición del Negocio

La ONE es una entidad creada para proponer, organizar y ejecutar, según corresponda, la aplicación de la política estatal en materia de estadística del país. Mediante el SEN, ejerce una adecuada dirección, ejecución y control de la captación de los indicadores económicos y sociales; así como su difusión de acuerdo con los requerimientos de la economía y necesidades del país, en términos de información estadística. Tiene como visión construir un sistema estadístico profesional, capaz de responder con calidad a las necesidades de información del país, para enfrentar los objetivos de desarrollo económico y social así como su adecuado reflejo internacional.

El proyecto SIGOB nace de la necesidad de centralizar toda la información existente en la ONE para lograr un mejor monitoreo y control de los datos estadísticos. Se enfoca en la creación de una herramienta que permita acceder a toda la información, con el objetivo de apoyar la toma de decisiones en las diferentes áreas socioeconómicas. Una de estas áreas es Protección del Trabajo y Seguridad Social, en la que se gestionan las estadísticas relacionadas con: la asistencia social, la seguridad social, la protección del trabajo y los pensionados.

La información estadística se desglosa por: territorios, organismos, sectores económicos y Consejos de Administración Provincial. Dicha información es recopilada con una periodicidad trimestral, semestral y anual, por los Centros Informantes que, posteriormente envían a la ONE Provincial, la cual los transfiere a la ONE Nacional, en las fechas definidas por esta última.

2.1.2 Temas de Análisis

Para desarrollar el Mercados de datos es necesario identificar los temas de análisis, cuyo objetivo es obtener varias perspectivas que orientan el avance del cumplimiento de las tareas planteadas y garantizar la utilidad y el éxito del diseño de las estructuras que se desarrollan.

La ONE manipula la información referente a los indicadores seleccionados de: asistencia social, seguridad social, protección del trabajo y pensionados, pertenecientes al área Protección del Trabajo y Seguridad Social. Es por ello que se han identificado a los indicadores seleccionados de **Protección del Trabajo y Seguridad Social**, como el principal tema de análisis.

2.1.3 Levantamiento de requisitos

Para el buen desarrollo del análisis en el proceso del negocio es preciso conocer qué es lo que necesitan los usuarios. La implicación de los mismos durante el ciclo de vida del producto es de gran importancia, ya que de ello depende que los resultados sean satisfactorios o no, en correspondencia a las necesidades de los usuarios. En la ONE, para el análisis de los indicadores de Protección del Trabajo y la Seguridad Social, los especialistas se enfocan en los indicadores seleccionados de: Protección del Trabajo, la Seguridad Social, la Asistencia Social y los Pensionados. Todo este análisis está destinado a satisfacer necesidades gubernamentales.

2.1.3.1 Requisitos de información

Los requisitos de información son todos los reportes que necesita visualizar el cliente. De acuerdo con la investigación del negocio se definieron 89 requisitos de información, de los cuales 20 corresponden a la Protección del Trabajo, 50 a la Seguridad Social, 14 a la Asistencia Social y cinco a los Pensionados. A continuación se relacionan los cinco requisitos de información definidos para los Pensionados y el resto se pueden apreciar en el **Anexo 4**.

RI1. Obtener la cantidad de pensionados beneficiarios de la Seguridad Social, de acuerdo al indicador Edad en Cuba, por años.

RI2. Obtener la cantidad de pensionados beneficiarios de la Seguridad Social, de acuerdo al indicador Invalidez Parcial en Cuba, por años.

RI3. Obtener la cantidad de pensionados beneficiarios de la Seguridad Social, de acuerdo al indicador Muerte en Cuba, por años.

RI4. Obtener la pensión media en Cuba, por años.

RI5. Obtener la cantidad de beneficiarios de la Seguridad Social vigentes en Cuba, por años.

2.1.3.2 Requisitos funcionales

Los requisitos funcionales son capacidades o condiciones que el sistema debe cumplir. Deben estar orientados a satisfacer las expectativas del cliente. A continuación se relacionan algunos de los requisitos funcionales que han sido identificados para desarrollar el Mercado de datos; el resto se pueden apreciar en el **Anexo 5**.

RF1 Autenticar usuario: el sistema debe permitir la autenticación del usuario y darle acceso a los reportes analíticos en el portal, así como mostrar los reportes disponibles para el usuario autenticado.

RF2 Crear reporte: El sistema debe tener disponibles la herramienta de creación de cubos multidimensionales, la herramienta de diseño de reportes multidimensionales, así como las estructuras del Mercado de datos, para la creación de los nuevos reportes.

RF3 Modificar Reporte: el sistema debe permitir que el usuario con permiso para modificar un reporte, ya autenticado en el sistema, pueda modificarlo.

RF4 Eliminar Reporte: el sistema debe permitir que el usuario con permiso para eliminar un reporte, ya autenticado en el sistema, pueda eliminarlo.

2.1.3.3 Requisitos no funcionales

Los requisitos no funcionales son las características que de una forma u otra pueden limitar el sistema, ejemplos de estos son la seguridad, el rendimiento, la fiabilidad entre otros. Son además propiedades o cualidades que el producto debe tener.

Para el desarrollo del Mercado de datos “Protección del Trabajo y Seguridad Social” se identifican 27 requisitos no funcionales, relacionados en el **Anexo 6**. De ellos: seis de usabilidad, cuatro de fiabilidad, uno de eficiencia, uno de soporte, cuatro de restricciones del diseño, uno para la documentación de usuarios y ayuda del sistema, cuatro de interfaz, tres de seguridad, uno de software y dos de hardware.

2.1.4 Perfilado de los datos

Es un proceso que consiste en analizar las fuentes de datos para conocer el estado, la calidad y la estructura en que se encuentran las mismas. A través del análisis de los resultados arrojados por este proceso, se infieren nuevas reglas del negocio, que posteriormente pasarán a ser las reglas de trasformación, llevadas a cabo en la implementación del subsistema de integración.

Al analizar los resultados obtenidos (para mayor precisión ver el **Anexo 7**), se concluye que:

- Los tipos de datos son varchar y enteros.
- No existen valores nulos o vacíos.
- No existen valores negativos.
- El valor mínimo encontrado es cero.
- La cantidad mínima de caracteres encontrados es uno.

2.1.5 Reglas del negocio

Las reglas del negocio describen las políticas, normas, operaciones, definiciones y restricciones presentes en una organización. Son de gran importancia para el logro de los objetivos en la misma.

En el proceso de almacenamiento de los indicadores de Protección del Trabajo y Seguridad Social se guarda la información del trabajo, la seguridad social, la asistencia social y de los pensionados en forma de clasificadores para un mejor control de estos parámetros. Varios de estos clasificadores tienen significados propios, que conllevan a las reglas del negocio. A continuación se relacionan las mismas:

Clasificador División Política Administrativa (DPA)

RN 1: Describe a los indicadores que están siendo analizados por provincias y municipios. La provincia posee un código de dos dígitos que la identifica y el municipio posee un código de cuatro dígitos, donde los dos primeros números corresponden al código de la provincia a la cual pertenece dicho municipio.

Clasificador de Actividad Económica (CAE)

RN 2: Divide los indicadores seleccionados por actividades económicas. Posee un código de seis dígitos que identifica la subrama económica, donde los cuatro primeros dígitos constituyen el código de la rama económica, de los cuales, los dos primeros dígitos constituyen el código del sector económico.

Nomenclador de Actividad Económica (NAE)

RN 3: Divide los indicadores seleccionados por actividades económicas. Posee un código de seis dígitos, donde los cuatro últimos dígitos constituyen el código de la clase y de ellos, los dos primeros constituyen el código de la división.

Otras reglas del Negocio:

RN 4: Una vez cargados los datos en el almacén, no pueden existir campos nulos.

RN 5: Los identificadores de las tablas no pueden tomar valores repetidos.

RN 6: Las medidas de las tablas de hechos deben poseer valores mayores o iguales a cero, nunca un número negativo.

2.1.6 Casos de uso del sistema

Los casos de uso del sistema representan información de manera visual. Describen lo que debe hacer el sistema relacionado con el usuario, representando generalmente requisitos funcionales. Para una mejor comprensión de los mismos se hace una representación gráfica que contiene los casos de uso y la relación que mantiene con los usuarios que interactúan con él. Para la realización del Mercado de datos fue necesario identificar casos de uso de información y funcionales.

Los casos de uso informativos se agrupan por el tipo de información que se maneja en la ONE específicamente en el área “Protección del Trabajo y Seguridad Social” y en dependencia de las necesidades de información de los usuarios. Estos se nombran:

- Analizar indicadores de Protección del Trabajo.
- Analizar indicadores de Seguridad Social.
- Analizar indicadores de Asistencia Social.
- Analizar indicadores de Pensionados.

Los casos de uso funcionales identificados están basados en la ejecución de las operaciones de ETL que se le realizarán a los clasificadores que contienen los datos estadísticos de Protección del Trabajo y Seguridad Social, además, se basan en la gestión y autenticación de usuarios, gestión de reportes para la aplicación y en la realización de las consultas en la Base de Datos. Estos se nombran:

- Autenticar usuario.
- Extraer datos de los ficheros fuentes.
- Realizar carga y transformación de los datos.
- Administrar reportes OLAP.
- Administrar roles.
- Administrar usuarios.
- Administrar permisos.

- Visualizar Reporte.

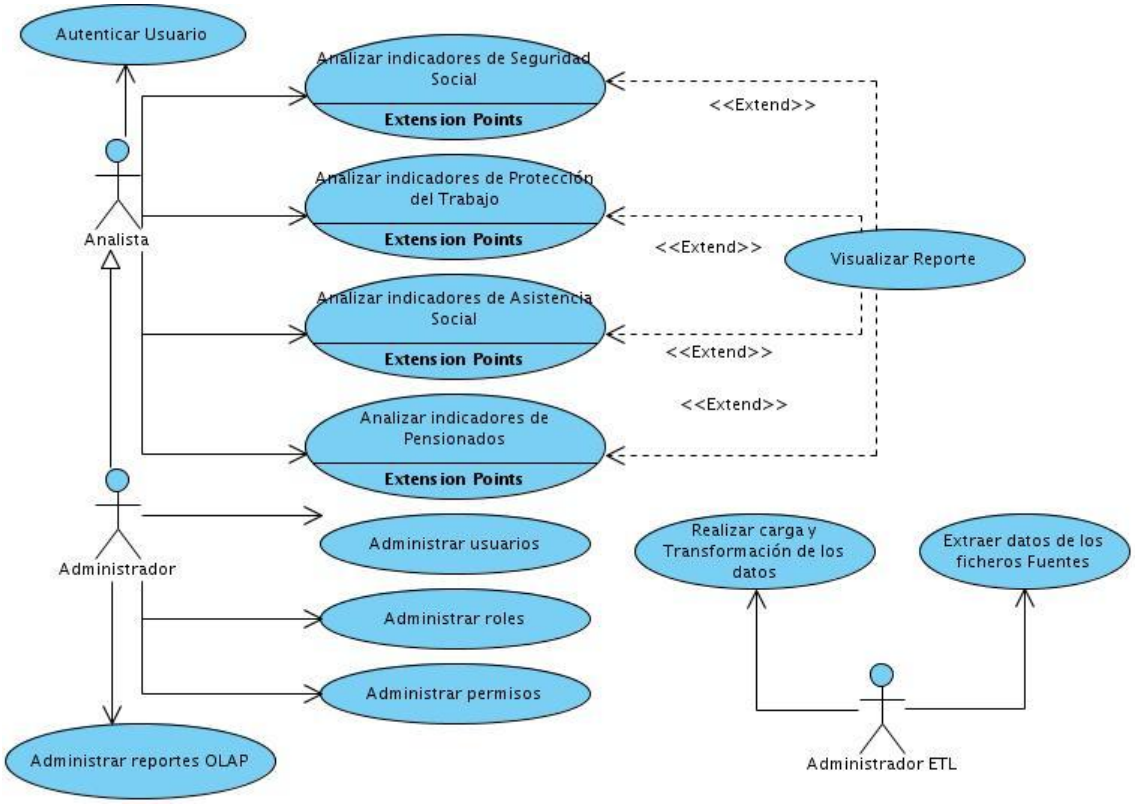


Ilustración 7: Diagrama de Caso de Uso.

Actores	**Descripción**
Analista	Es el responsable de analizar la información de los diferentes indicadores de Protección del Trabajo y Seguridad Social.
Administrador	Es el responsable de la gestión de los usuarios y permisos, así como la administración de los reportes OLAP.
Administrador ETL	Es el responsable de la extracción, transformación y carga de los datos.

Tabla 2: Actores y Descripciones.

Caso de Uso	Descripción
Analizar indicador Protección del Trabajo	Visualiza los reportes de los indicadores seleccionados de la Protección Trabajo.
Analizar indicador de Seguridad Social	Visualiza los reportes de los indicadores seleccionados de la Seguridad Social.
Analizar indicador de la Asistencia Social	Visualiza los reportes de los indicadores seleccionados de la Asistencia Social.
Analizar indicadores de los Pensionados	Visualiza los reportes de los indicadores seleccionados de los Pensionados. Para más información ver el **Anexo 8**.
Realizar carga y transformación de los datos	Realiza la transformación y carga de los datos necesarios para la construcción del Mercado de datos.
Extraer datos de los ficheros fuentes	Realiza la extracción de los datos necesarios de los ficheros fuentes DBF y XLS.
Administrar roles	Crea o elimina los roles y asigna o rehúsa usuarios a los roles.
Administrar usuarios	Crea o elimina los usuarios que interactúan con el sistema y asigna o rehúsa roles a los usuarios.
Administrar reporte OLAP	Elimina, inserta o modifica los reportes OLAP que se visualizan.
Administrar permiso	Asigna, modifica o deniega los permisos que existan sobre un rol o un usuario en el sistema.
Autenticar usuario	Realiza la autenticación de los usuarios en el sistema. Si esto no se logra no podrá ejecutarse ningún otro caso de uso. Para más información ver el **Anexo 8**.
Visualizar Reporte	Realiza la visualización de los reportes y las

	operaciones o funciones deseadas.

Tabla 3: Casos de Uso y descripciones.

2.1.7 Definición de los reportes candidatos

Con el objetivo de responder a las necesidades de información, se realiza una propuesto de 13 reportes candidatos, identificados en el área de Protección del Trabajo y Seguridad Social. Los mismos han sido agrupados por Libros de Trabajo (L.T), en dependencia de los requisitos de información y estos están contenidos dentro del Área de Análisis (A.A) Protección del Trabajo y Seguridad Social. A continuación se relacionan los reportes identificados por L.T.

LT Protección del Trabajo: Contiene cinco reportes que permiten realizar un análisis general de datos en función de los indicadores: cantidad de incidentes, cantidad de accidentes fatales, cantidad de lesionados por accidentes del trabajo y los días perdidos por accidentes del trabajo. Los reportes son:

- TS1 – Protección del Trabajo. Indicadores seleccionados en Cuba.
- TS2 – Protección del Trabajo. Indicadores seleccionados por Consejos de Administración Provincial.
- TS3 – Protección del Trabajo. Indicadores seleccionados por organismos.
- TS4 – Protección del Trabajo. Indicadores seleccionados por provincias.
- TS5 – Protección del Trabajo. Indicadores seleccionados por sectores económicos.

LT Seguridad Social: Contiene cinco reportes que permiten realizar un análisis general de datos en función de los indicadores: trabajadores subvencionados por enfermedad y accidente común, trabajadores subvencionados por enfermedad profesional y accidentes del trabajo, trabajadoras acogidas a la licencia de maternidad, trabajadores reubicados por invalidez parcial, trabajadores acogidos a la Resolución No. 22/03, subvenciones pagadas por enfermedad y accidente común, subvenciones pagadas por enfermedad profesional y accidentes del trabajo, pagado a trabajadoras acogidas a la licencia de maternidad, prestación social pagada a trabajadores acogidos a la Resolución No. 22/03 y el pagado a reubicados por invalidez parcial. Los reportes son:

- TS1 – Seguridad Social. Indicadores seleccionados en Cuba.
- TS2 – Seguridad Social. Indicadores seleccionados por Consejos de Administración Provincial.
- TS3 – Seguridad Social. Indicadores seleccionados por organismos.
- TS4 – Seguridad Social. Indicadores seleccionados por provincias.
- TS5 – Seguridad Social. Indicadores seleccionados por sectores económicos.

LT Asistencia Social: Contiene dos reportes que permiten realizar un análisis general de datos en función de los indicadores: gastos por la asistencia social, beneficiarios de la asistencia social, núcleos protegidos por la asistencia social, adultos mayores beneficiarios de la asistencia social, personas con discapacidad beneficiaria de la asistencia social, madres de hijos con discapacidad severa beneficiarias de la asistencia social y los beneficiarios del servicio de asistente social a domicilio. Los reportes son:

- TS1 – Asistencia Social. Indicadores seleccionados en Cuba.
- TS2 – Asistencia Social. Indicadores seleccionados por provincias.

LT Pensionados: Contiene un reporte que permiten realizar un análisis general de datos en función de los indicadores: pensionados por Edad, pensionados por Invalidez Total, pensionados por Muerte, cantidad de beneficiarios vigentes, y la pensión media. El reporte identificado es:

- TS1 – Pensionados. Indicadores seleccionados en Cuba.

En la **Tabla 4** se muestra la descripción del reporte candidato contenido dentro del L.T Pensionados.

Área de análisis (A.A)	Protección del Trabajo y Seguridad Social
Libro de Trabajo (L.T)	L.T Pensionados
Reporte (Tabla de Salida – TS)	TS1 – Pensionados. Indicadores seleccionados en Cuba
Descripción	Reporte que muestra los indicadores de los Pensionados en Cuba
Elementos del reporte	Indicador Tiempo
Frecuencia de emisión	Anual
Gráfico	Gráfico de barras 3D (tres dimensiones)

Tabla 4: Descripción del reporte “TS1 - Pensionados. Indicadores seleccionados en Cuba”.

2.2 Diseño del Mercado de datos

Para diseñar el Mercado de datos “Protección del Trabajo y Seguridad Social”, es necesario realizar el diseño de los tres subsistemas que lo componen.

2.2.1 Diseño del subsistema de almacenamiento

Aspectos como la identificación del gránulo del proceso, las dimensiones, las tablas de hechos y la política de recuperación y respaldo, constituyen elementos esenciales para realizar el diseño de un subsistema de almacenamiento.

2.2.1.1 Granularidad del proceso

La granularidad, como concepto, es una medida del nivel de detalle enfocada a cada ocurrencia que exista en la tabla de hechos. Por esta razón se puede inferir la estrecha relación existente entre las dimensiones y la granularidad (7), es decir, el gránulo es equivalente a los niveles jerárquicos que poseen las dimensiones. La elección de la granularidad depende en gran medida de los requerimientos del negocio.

Mientras mayor sea el nivel de detalle de los datos, se tendrán mayores posibilidades analíticas, ya que los mismos podrán ser resumidos o sumarizados. Es decir, los datos que posean granularidad fina (nivel de detalle) podrán ser resumidos hasta obtener una granularidad media o gruesa. No sucede lo mismo en sentido contrario, ya que los datos almacenados con granularidad media podrán resumirse, pero no tendrán la facultad de ser analizados a un mayor nivel de detalle.

Conociendo que los datos referentes a la Protección del Trabajo poseen una periodicidad trimestral y los datos pertenecientes a la Seguridad Social tienen una periodicidad semestral, por ende, la granularidad con que se guardan los registros de la Protección del Trabajo es a nivel de trimestre, donde estos datos podrán sumarizarse por semestres y años, en cambio, los registros de la Seguridad Social que se almacenan a nivel de semestre, solo podrán sumarizarse por años.

En concordancia con las necesidades del negocio el grano queda definido como la información estadística trimestral perteneciente a todos los Centros Informantes, registrados bajo la estructura de organismos, Clasificador de Actividades Económicas y Nomenclador de Actividades Económicas, en todos los municipios, de los indicadores estadísticos captados en el modelo M5201 del SIEN.

2.2.1.2 Dimensiones

Después de haber declarado el grano del proceso a modelar se comienza la definición de las dimensiones candidatas que posteriormente, después de un profundo análisis, se convertirán en las dimensiones que contendrá la solución. Las dimensiones poseen entre sus características principales la definición de jerarquías entre sus atributos las que poseen como objetivo plasmar explícitamente la forma en que se puede consolidar la realización del proceso de análisis en línea de la información, ya sea mediante el uso de sumas, porcientos, máximos, mínimos, etc.

A continuación se describen las dimensiones y jerarquías que están relacionadas con el repositorio principal donde se va a almacenar la información atómicamente.

Dimensión CAE (dim_cae)

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información atendiendo al Clasificador de Actividades Económicas.

Jerarquía:

- Sector -> Rama ->CAE

Dimensión DPA (dim_dpa)

Esta dimensión almacena los valores pertenecientes a la División Política Administrativa que presenta el país.

Jerarquía:

- Provincia -> Municipio

Dimensión NAE (dim_nae)

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información atendiendo al Nomenclador de Actividades Económicas.

Jerarquía:

- NAE
- Sección -> División -> Clase

Dimensión Organismo (dim_organismo)

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información atendiendo al Organismo al cual pertenece el Centro Informante que suministra la información.

Jerarquía:

- Organismo

Dimensión Provincia (dim_provincia)

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información atendiendo a la provincia bajo la cual se subordina el Centro Informante que suministra la información.

Jerarquía:

- Provincia

Dimensión Indicadores Generales (dim_indicador_general)

Describe un valor mediante el cual puede clasificarse la información de la empresa, definiendo los indicadores.

Jerarquía:

- Indicador

Dimensiones Temporales

Estas dimensiones son las más comunes e importantes en el diseño de Mercados de datos debido a que define una línea de tiempo para enmarcar la información almacenada y organiza jerárquicamente cuando fue captada la información. En el Mercado de datos “Protección del Trabajo y Seguridad Social” se tienen tres dimensiones temporales.

- **Dimensión Temporal Año (dim_temporal_anno)**

  Jerarquía:

  - Año

- **Dimensión Temporal Semestre (dim_temporal_semestre)**

  Jerarquía:

  - Año -> Semestre

- **Dimensión Temporal Trimestre (dim_temporal_trimestre)**

  Jerarquía:

  - Año -> Semestre -> Trimestre

2.2.1.3 Hechos

Las tablas de hechos son las que almacenan las medidas numéricas. Para el Mercado de datos “Protección del Trabajo y Seguridad Social” se identifican cuatro hechos:

- **Hecho Protección del Trabajo (hech_proteccion_trabajo).** Para esta tabla de hechos se definen como medidas numéricas los dos valores que se captan en el modelo M5201

concernientes al valor real del año actual y del año anterior, debido a que es un modelo estadístico netamente económico.

- **Hecho Seguridad Social (hech_seguridad_social).** En este caso se definen como medidas numéricas los dos valores que se captan en el modelo M5201 concernientes al valor real del año actual y del año anterior, debido a que es un modelo estadístico netamente económico.
- **Hecho Asistencia Social (hech_asistencia_social).** En este caso se definen como medida numérica el valor que se capta en la serie 7.16 “Indicadores seleccionados de la Asistencia Social” concerniente al valor real del año actual.
- **Hecho Pensionados (hech_pensionados).** En este caso se definen como medida numérica el valor que se capta en la serie 7.14 “Indicadores seleccionados de los Pensionados” concerniente al valor real del año actual.

2.2.1.4 Matriz BUS o Dimensional

La Matriz Dimensional es la representación de la relación que existe entre los hechos y las dimensiones. Se define como la habilidad para describir y seguir la vida tanto de una dimensión como de un hecho, la cual permite determinar el impacto que provocaría un cambio durante el desarrollo del sistema.

Dimensiones:

- D1: dim_nae
- D2: dim_cae
- D3: dim_dpa
- D4: dim_organismo
- D5: dim_provincia
- D6: dim_indicador_general
- D7: dim_temporal_anno
- D8: dim_temporal_semestre
- D9: dim_temporal_trimestre

Hechos:

- H1: hech_proteccion_trabajo
- H2: hech_seguridad_social
- H3: hech_asistencia_social
- H4: hech_pensionados

Hechos/Dimensiones	D1	D2	D3	D4	D5	D6	D7	D8	D9
H1	x	x	x	x		x			x

H2	x	x	x	x		x		x	
H3					x	x	x		
H4			x			x	x		

Tabla 5: Matriz BUS.

2.2.1.5 Modelo físico de los datos

Una vez definido dentro del negocio las dimensiones, medidas y la granularidad, se procede a la estructuración del modelo o los modelos físicos que existirán. En tal sentido se puede destacar que, por las necesidades actuales del negocio, existen varias tablas de hechos que unifican las dimensiones definidas y las medidas que se han especificado hasta el momento. A continuación se muestra una porción del modelo físico de los datos y en el **Anexo 9** se puede apreciar en su totalidad.

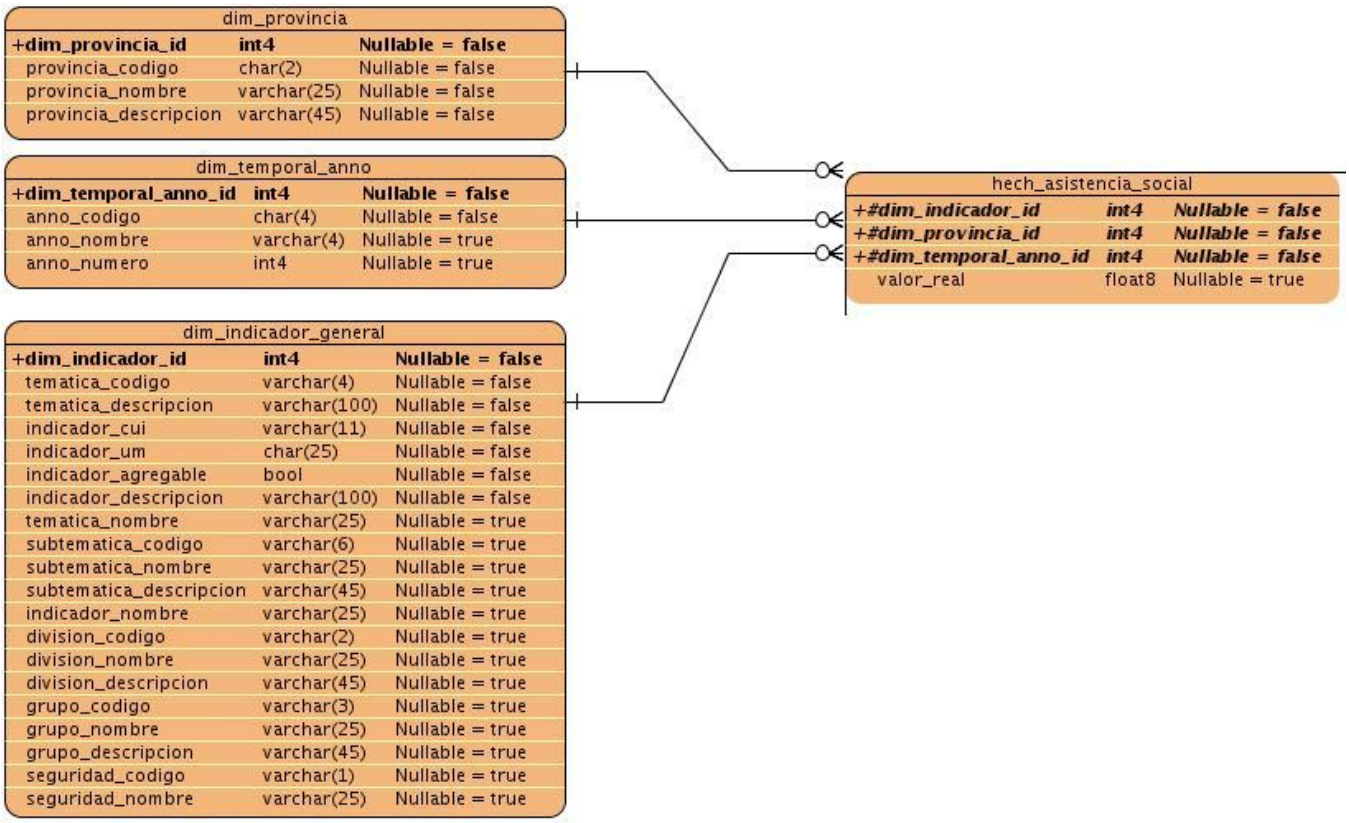


Ilustración 8: Porción del modelo físico de los datos.

2.2.1.6 Política de recuperación y respaldo

La política de respaldo y recuperación que se empleará en el Mercado de datos está condicionada por tres puntos fundamentales:

- **Periodicidad de salvas del sistema:** estas se harán constantemente, a toda la información contenida en el Mercado de datos, en un período aproximado de 90 días y verificando siempre la existencia de una copia de toda la información almacenada.
- **Tablas involucradas:** las tablas que se involucran en la realización son: hech_proteccion_trabajo, hech_seguridad_social, hech_asistencia_social y hech_pensionados.
- **Salvas existentes:** actualmente no existen salvas en esta área, por lo que se prevé la realización de reemplazos cada 90 días y un chequeo trimestral de su estado, mediante pruebas de rendimiento y flexibilidad.

Estrategia de Copias de Respaldo

Para garantizar la persistencia de la información y la contribución a no tener almacenada información que no sea útil a los analistas estadísticos de la ONE se definieron las siguientes directrices.

Se realizarán backups con periodicidad trimestral, de la información total que posea la Base de Datos garantizando en todo momento que exista una copia exacta de la información que está vigente en el servidor. Se realizará en periodos trimestrales debido a que la ONE tiene definido que este modelo estadístico se capte a las entidades de esta manera. La estructura de carpetas definidas con este sentido será con la jerarquía Modelo -> Año -> Semestre -> Trimestre y el nombre del scripts será de la misma forma anno_modelo_semestre_trimestre, logrando dejar plasmado claramente la fecha de la copia realizada.

2.2.2 Diseño del subsistema de Integración

El diseño de las transformaciones constituye un elemento esencial para lograr el diseño del subsistema de integración.

2.2.2.1 Diseño de los procesos de Integración

Una vez analizados los datos se procede a realizar el diseño de las transformaciones. De estos diseños a la implementación de las transformaciones suelen haber variaciones, debido a que en el proceso de elaboración de las transformaciones suelen aparecer situaciones con los datos y estrategias para resolver estas situaciones, las cuales constituyen reglas de transformación.

Para cargar los indicadores del Mercado de datos “Protección del Trabajo y Seguridad Social” se diseña la siguiente transformación.

Ilustración 9: Diseño de la transformación para la carga de los indicadores.

Para cargar los hechos del Mercado de datos “Protección del Trabajo y Seguridad Social” se diseñan dos transformaciones, donde el primer diseño corresponde a las transformaciones para los hechos Asistencia Social y Pensionados.

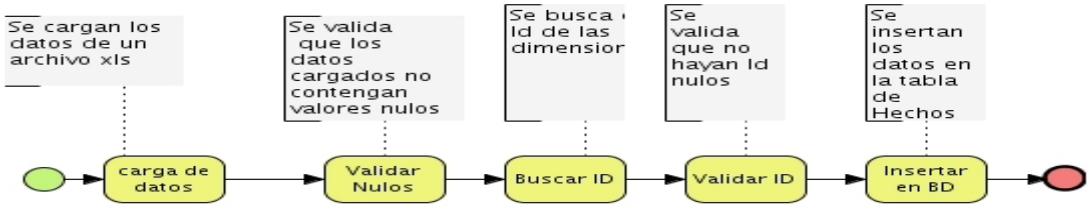


Ilustración 10: Diseño de transformación para carga de los datos de las tablas de hechos hech_asistencia_social y hech_pensionados.

El segundo diseño corresponde a la transformación para los hechos Protección del Trabajo y Seguridad Social.

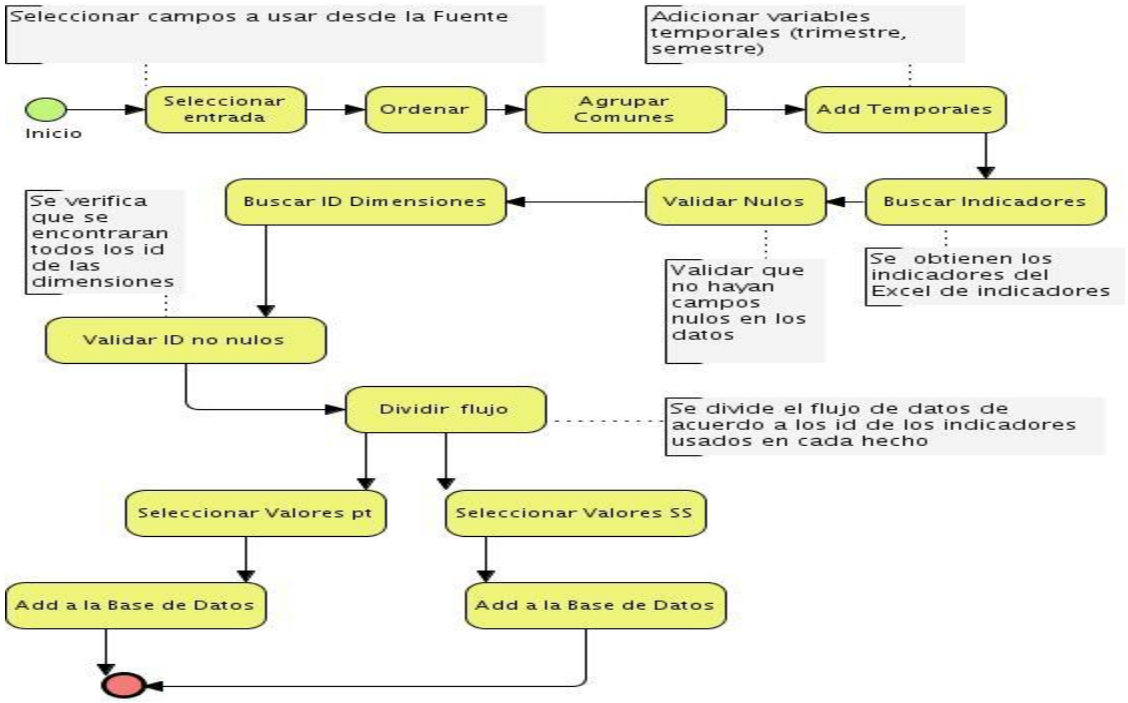


Ilustración 11: Diseño de la transformación para la carga de los datos de las tablas de hechos hech_proteccion_trabajo y hech_seguridad_social.

2.2.3 Diseño del subsistema de visualización

Existen diversas estructuras de datos, a través de las cuales se puede representar la información contenida en un Mercado de datos, de las cuales, los cubos multidimensionales constituyen una de las estructuras más utilizadas.

2.2.3.1 Diseño de los Cubos OLAP

Un cubo multidimensional o hipercubo representa o convierte los datos planos que se encuentran en filas y columnas de una tabla, razón por la cual en el Mercado de datos “Protección del Trabajo y Seguridad Social” se tienen cuatro cubos multidimensionales, o sea, un cubo por cada tabla de hechos y nueve dimensiones.

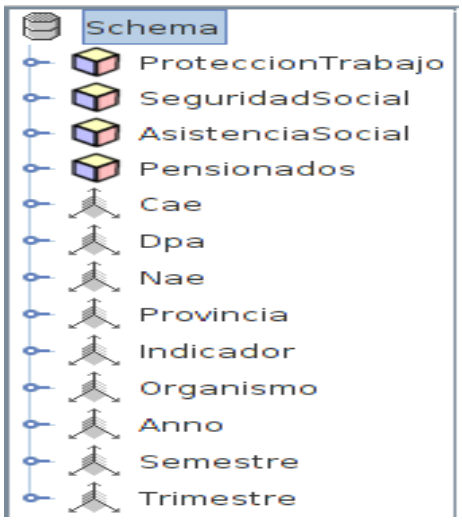


Ilustración 12: Esquema, cubos y dimensiones del Mercado de datos.

El cubo Protección del Trabajo contiene cinco dimensiones usables y dos medidas.

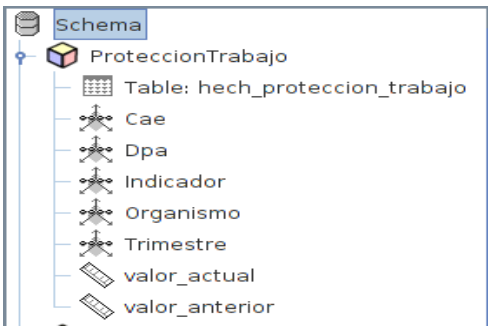


Ilustración 13: Cubo "Protección del Trabajo".

El cubo Seguridad Social contiene cinco dimensiones usables y dos medidas.

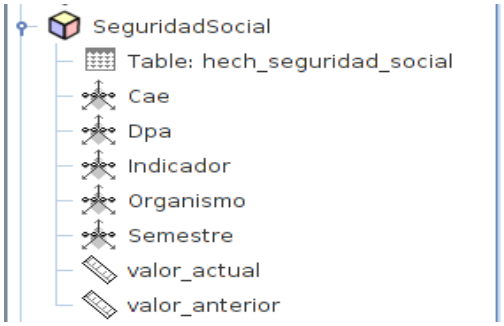


Ilustración 14: Cubo "Seguridad Social".

El cubo Asistencia Social contiene tres dimensiones usables y una medida.

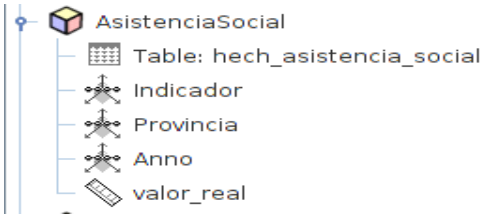


Ilustración 15: Cubo "Asistencia Social".

El cubo Pensionados contiene dos dimensiones usables y una medida.

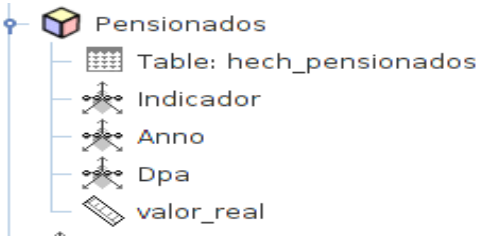


Ilustración 16: Cubo "Pensionados".

2.2.4 Esquema de seguridad

La seguridad para el Mercado de datos “Protección del Trabajo y Seguridad Social” está determinada mayormente por los niveles de acceso al sistema. Dicha seguridad se rige fundamentalmente por los roles y permisos que los usuarios poseen en su interacción con la Base de Datos y la aplicación.

2.2.4.1 Seguridad en la Base de Datos

Para la interacción de los usuarios con la Base de Datos se definen los siguientes actores:

Actor	Permiso
Administrador ETL	Tiene acceso a las tablas de la Base de Datos del sistema, con el propósito de realizar el proceso de ETL.
Administrador	Tiene total acceso a la Base de Datos del sistema.
Analista	Tiene acceso de lectura a la Base de Datos del sistema.

Tabla 6: Actores y Permisos.

2.2.4.2 Seguridad en la aplicación

Actualmente las aplicaciones desplegadas en el Servidor de Inteligencia de Negocios de Pentaho muestran un sucesivo incremento, así como los usuarios que tiene acceso a estas. Como consecuencia de esto se definen los siguientes roles:

Roles	Permisos
administrador	• Gestiona “Sistema de Información de Gobierno”. • Tiene acceso total a todas las A.A, libros de trabajos y reportes. • Visualiza los reportes
análisis	• Tiene acceso de solo lectura al A.A Protección del Trabajo y Seguridad Social, incluyendo los libros de trabajos y reportes. • Visualiza los reportes.

Tabla 7: Roles y permisos.

Elemento de aplicación	Roles con acceso
A.A.G SIGOB (Área de Análisis General del Sistema de Información de Gobierno)	• administrador • análisis
Carpeta raíz: A.A Protección del Trabajo y Seguridad Social.	• administrador • análisis

Tabla 8: Elementos de aplicación y roles con acceso.

Además de los niveles de accesos y roles antes definidos para interactuar con la aplicación, la Plataforma Pentaho BI tiene incluida su propia seguridad, la cual está basada en la infraestructura proporcionada por el Sistema de Seguridad Acegi. El mismo se divide en cuatro áreas fundamentales:

- Seguridad de acceso a datos de objetos: Incluye usuarios, contraseñas, autorizaciones permitidas, recursos web y protección a datos.
- Autenticación: Tiene que ver con el procesamiento de información interactiva de inicio de sesión (por ejemplo nombre de usuario y contraseña).
- Autorización de recursos web (URL): Brinda protección a las URL para responder a cada usuario si pueden o no acceder a una determinada página. Esto es decidido por el administrador de recursos web, el cual le brinda a cada usuario autenticado un permiso de seguridad, delimitando las páginas a las que tiene acceso y a las que no.
- Autorización a objetos del dominio: En el sistema, los únicos objetos del dominio protegidos por la plataforma son los objetos de repositorio otorgados al usuario autentificado. Es responsabilidad de los objetos del dominio autorizar las operaciones solicitadas por este.

Conclusiones del capítulo

Una vez finalizado el capítulo y con ello el análisis y diseño del Mercado de datos “Protección del Trabajo y Seguridad Social”, se puede concluir que:

- Se refinaron las principales reglas del negocio que serán empleadas en la implementación del Mercado de datos cubriendo las peticiones y necesidades del cliente.
- Se realizó la especificación y el diseño de los casos de uso del sistema, de los cuales, cuatro son de información y ocho funcionales, posibilitando que el sistema cumpla con las funciones y requerimientos necesarios.
- Se diseñó el modelo de datos, identificando nueve tablas dimensionales y cuatro tablas de hechos que garantizan la integridad de los datos y un buen funcionamiento del sistema.
- Se definieron las políticas de recuperación y respaldo de la información, garantizando la fiabilidad, conservación y persistencia de los datos.
- La identificación de los reportes candidatos y el diseño de los cubos multidimensionales satisfacen las necesidades de información del cliente.

Capítulo 3. Implementación del Mercado de datos

Introducción

Según la metodología que se está utilizando posterior al diseño dimensional del Mercado de datos se procede a la implementación. Para ello es necesario implementar los subsistemas de almacenamiento, integración y visualización.

3.1 Implementación del subsistema de almacenamiento

Para implementar el subsistema de almacenamiento es de vital importancia estandarizar los nombres, desarrollar el modelo físico de los datos, plantear la estrategia de indexado, construir una instancia de la Base de Datos y desarrollar la estructura física de almacenamiento.

3.1.1 Estandarización de los Nombres

El objetivo principal de este paso es organizar la forma en que se van a denominar las estructuras con el fin de que quede documentado para su utilización. Además es conveniente mantener una nomenclatura estándar en el nombrado para un mejor entendimiento de las estructuras por los desarrolladores.

En la solución propuesta, a nivel global (ver **Anexo 10** “Estandarización de los nombres”), se mantuvo la misma estructura en cuanto a la clasificación, específicamente en lo referente a si la estructura es una dimensión o una tabla de hecho. Si la tabla es una dimensión al nombre le precede las letras “dim” ejemplo dim_dpa, en caso de ser una tabla de hecho se le antepone las siglas “hech”, ejemplo, hech_pensionados.

En el caso de los atributos de las dimensiones se siguió la misma política en todas. Cuando se refiere a la llave de las dimensiones se le denominó “dim_dimension_id”, en caso que fuera algún código del negocio se le especificó “dimension_codigo”. Así mismo con respecto a las descripciones: “dimension_descripcion”. A modo de generalizar se puede decir que todos los atributos se nombraron como “dimension_clase”, excepto cuando la dimensión posee más de un nivel jerárquico, en este caso se siguió la estrategia de nombrarlos “nombre_de_la_jerarquía_clase”, ejemplo de esto es en la dimensión dim_indicador_general el atributo “tematica_descripcion”.

Con respecto a las medidas se nombraron mediante la agrupación de la clase “valor”, los clasificadores “actual”, “real”, “anterior”. Esta estructura es semejante a su nombre especificado dentro del modelo estadístico en cuestión. Un ejemplo de esto es la medida “valor_actual” que dentro del modelo aparece como Año Actual.

Al finalizar este paso queda completamente estructurado la nomenclatura utilizada para la denominación de las tablas, atributos y medidas dentro de la Base de Datos. Ya después de tenerlos definidos es que se comienza con la implementación de las estructuras físicas.

3.1.2 Estrategia de Indexado

Sobre un Mercado de datos se realizarán, muchas veces, consultas de gran complejidad que solicitarán información que cumpla determinados criterios, es decir, los usuarios frecuentemente querrán especificar los valores con los cuales se filtrarán los datos que deberán ser retornados. La mayoría de estas consultas incluirán, probablemente, operaciones de join entre tablas muy grandes, lo cual puede resultar extremadamente costoso. Para ganar en eficiencia a la hora de realizar estas operaciones se han investigado y creado técnicas especializadas que hoy ofrecen varios gestores, como los índices.

Un índice es una estructura física que permite un tipo de acceso alternativo al secuencial. Es creado a partir de una o varias columnas de una tabla, y, por lo general, es construido en forma de árbol balanceado (B-Tree). Al ser estructuras físicas, los índices van a tener un fichero asociado, en cuyas páginas se pueden almacenar uno o varios nodos del árbol. Cada uno de ellos apunta hacia otros nodos del árbol o hace referencia a las filas de la tabla. En cada nodo, los valores están ordenados, y los que se encuentran en un nodo hijo son menores o iguales que el valor en el nodo padre que le hace referencia. Los nodos que apuntan hacia las filas reciben el nombre de “páginas hojas”, y están enlazados entre sí: una página hoja apunta a otra hoja que contiene el próximo conjunto de valores. (7)

La solución más apropiada es crear índices para las llaves primarias y foráneas: debido a que las operaciones de join consumen mucho tiempo, y para la mayoría de ellos las columnas por las que se realiza la unión son llaves foráneas, crear índices en las llaves implicadas en la unión puede ser ventajoso. Por esta razón, el indexado que se utiliza es el que trae por defecto el gestor PostgreSQL, para la búsqueda de datos utilizando las llaves primarias y foráneas. Todas las llaves primarias, que son llaves subrogadas además, poseen índices de tipo “b-tree” (Árboles-B) lo que implica que cualquier búsqueda que se realice utilizando las llaves se optimizará mediante este método. Para las dimensiones el indexado propuesto es sobre la llave primaria de cada una de ellas, al igual que las tablas de hechos.

3.1.3 Desarrollo de la estructura física de almacenamiento

Existe un nivel que se sitúa por debajo de las estructuras de datos en el cual se encuentran los archivos, discos, particiones, espacios de tablas, etc. La utilización adecuada de estos elementos y el dominio de los mismos inciden significativamente en el éxito del Mercado de datos. (7)

Los elementos a tener en cuenta durante el desarrollo son el particionamiento de las tablas, en función de lograr una mayor organización de la información y velocidad en su recuperación, y estructuras de control de cambios con el fin de minimizar la utilización de recursos físicos cuando se refresquen las tablas de hechos.

3.1.3.1 Esquemas

Para garantizar este fin se definieron tres esquemas para almacenar las tablas:

- **dimensiones**: donde se ubican las dimensiones identificadas del Mercado de datos “Protección del Trabajo y Seguridad Social”.
- **mart_trab_seg_social:** donde se ubican las tablas de hechos del Mercado de datos “Protección del Trabajo y Seguridad Social” y los triggers o disparadores correspondientes a las tablas de hechos para realizar el control de los cambios.
- **metadatos:** donde se ubican las tablas, las funciones y las funciones triggers necesarias para realizar el control de cambios sobre las tablas de hechos.

3.1.3.2 Tablespace

Además de los esquemas definidos se utilizan un conjunto de tablespace para separar la utilización de los recursos físicos. Los tablespace utilizados son: **tb_medida** (para agrupar todas la tablas de hechos), **tb_dimension** (para todas las tablas de dimensiones) y **tb_indices** (para que sea utilizado por los índices b-tree creados).

3.1.3.3 Control de Cambios

La estructura diseñada para el control de cambios está basada específicamente en cinco tablas de metadatos con el fin de almacenar la acción que se realizó, sobre que tabla de hechos y los valores asociados. Las tablas en cuestión son: **tb_control_cambios**, **tb_cc_pens**, **tb_cc_asocial**, **tb_cc_ssocial** y **tb_cc_ptrab**. Las cuatro últimas, son una especialización de la primera con el fin de almacenar todos los datos de la tupla insertada, modificada o eliminada. En el **Anexo 11** se puede

apreciar el diseño de las tablas para el control de cambios y en el **Anexo 12** las funciones implementadas para controlar los cambios realizados sobre la tabla de hechos **hech_pensionados**.

3.1.4 Usuarios, Roles y Privilegios

PostgreSQL almacena los datos de usuarios así como también los datos de los grupos dentro de sus propios catálogos de sistema. De esta manera, cualquier conexión a PostgreSQL debe ser realizada con un usuario específico, y cualquier usuario puede pertenecer a uno o más grupos definidos. (16)

La tabla de usuarios en PostgreSQL controla los permisos de acceso y quién está autorizado a realizar acciones en el sistema, al igual que las acciones puede realizar. Los grupos existen como un mecanismo para simplificar la ubicación de estos permisos. Tanto las tablas de usuarios como de grupos existen como objetos globales de Base de Datos, por consiguiente no están agregadas a ninguna base de datos en particular. (16)

PostgreSQL incorpora tres niveles de acceso: nivel de Base de Datos, nivel de esquemas, nivel de tablas. Una Base de Datos puede abarcar varios esquemas diferentes y estos, a su vez contienen varias tablas. En el **Anexo 13** se relacionan los niveles de acceso y los privilegios asociados.

Para la Base de Datos perteneciente al Mercado de datos “Protección del Trabajo y Seguridad Social” se definen tres grupos en dependencia del nivel de acceso y las funciones que realizan:

- **rol_admin:** posee acceso total a la Base de Datos, tanto de administración como configuración de la Base de Datos, usuarios y roles.
- **rol_etl:** este rol se basa en seleccionar, insertar, actualizar y eliminar datos de las tablas existentes en la Base de Datos.
- **rol_analisis:** su rol se basa en consultar la información existente en la Base de Datos.

Los usuarios definidos son:

- **admin:** es superusuario y el propietario de la Base de Datos. Desempeña el rol **rol_admin**. Posee todos los privilegios en todos los niveles.
- **etl:** desempeña el rol **rol_etl**. En el nivel de Base de Datos solo posee privilegio “**c**” (connect), a nivel de esquemas “**U**” (Usage) y a nivel de tablas “**arwxdt**” (insert, select, update, references, delete, triggers).
- **analista:** desempeña el rol **rol_analisis**. En el nivel de Base de Datos solo posee privilegio “**c**” (connect), a nivel de esquemas “**U**” y en el de tablas “**rx**”.

3.2 Implementación del subsistema de integración

Las fuentes de datos constituyen ficheros dbf y xls. Para iniciar el proceso de extracción es necesario estandarizar los ficheros xls.

Posterior a la extracción de los datos el sistema se encuentra listo para realizar la etapa de transformación.

3.2.1 Implementación de las de transformaciones

Las transformaciones constituyen un elemento básico dentro de la implementación del proceso de ETL. Una transformación está compuesta por pasos, que constituyen el elemento más pequeño de la transformación y se encuentran unidos a través de saltos.

Para implementar el proceso de ETL del Mercado de datos “Protección del Trabajo y Seguridad Social” se realizan varias de transformaciones (ver **Anexo 14**).

carga_indicadores_generales: En esta transformación se cargan los indicadores del modelo M5201 del SIEN y de las series utilizadas, a partir de un fichero xls, hacia la tabla dim_indicador_general que se encuentra en la Base de Datos.

Para cargar los datos de las tablas de hechos se realizan tres transformaciones, donde una posee como entrada de datos los ficheros dbf y el resto los ficheros xls. Estas incluyen validaciones para que no se inserten datos nulos y para que existan las relaciones de los identificadores de cada dimensión con las tablas de hechos asociadas. Es de vital importancia destacar que los datos erróneos son salvados en ficheros xls para facilitar su manipulación en la corrección del error.

Para cargar los datos de las tablas de hechos **hech_proteccion_trabajo** y **hech_seguridad_social** se realiza la transformación **carga_hech_ptss**. La misma carga la información contenida en el modelo M5201 del SIEN correspondiente a los meses de marzo, junio, septiembre y diciembre.

Para cargar los datos de las tablas de hechos **hech_asistencia_social** se realiza una transformación, la cual posee como entrada de datos un fichero xls. **carga_hech_aistencia_social:** En esta transformación se carga la información contenida en la serie 7.16.

Para cargar los datos de las tablas de hechos **hech_pensionados** se realiza una transformación, la cual posee como entrada de datos un fichero xls. **carga_hech_pensionados:** En esta transformación se carga la información contenida en la serie 7.14.

3.2.2 Implementación de los trabajos

Un trabajo o job es un conjunto de tareas que tienen como objetivo realizar una acción determinada. En los trabajos se utilizan pasos específicos que son distintos a los disponibles en las transformaciones. Además, los jobs permiten ejecutar una o varias transformaciones que han sido diseñadas, siguiendo una secuencia de ejecución para cada elemento que los conforman. Mediante los trabajos se definen el horario y frecuencia de la carga, así como el orden en que van a ser ejecutadas las transformaciones, para poder realizar exitosamente la carga de los datos.

El trabajo correspondiente al Mercado de datos “Protección del Trabajo y Seguridad Social” es el siguiente:

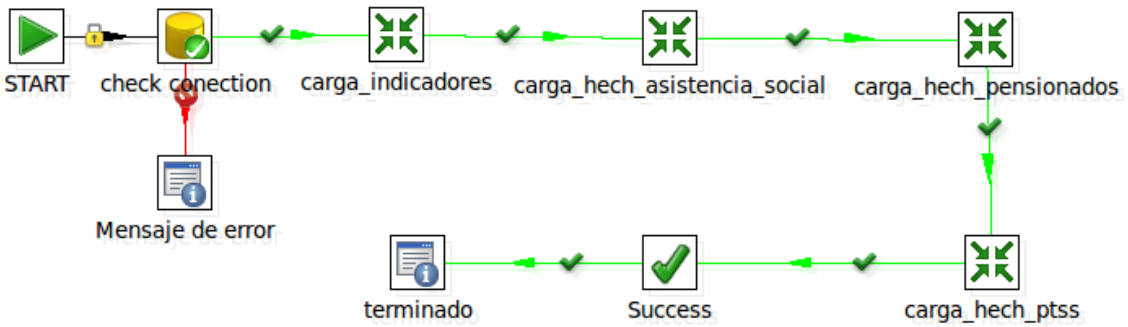


Ilustración 17: Trabajo o Job del Mercado de datos "Protección del Trabajo y Seguridad Social".

3.3 Implementación del subsistema de visualización

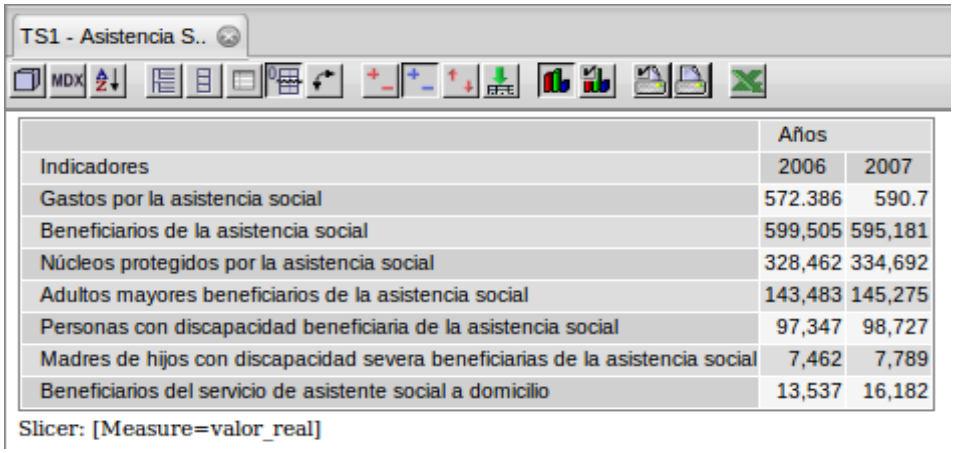
Una vez realizado el diseño de los reportes candidatos, el diseño de los cubos OLAP y la carga de los datos en el Mercado de datos, se procede a implementar los reportes candidatos, con el objetivo de completar el subsistema de visualización.

3.3.1 Implementación de los reportes candidatos

Las Expresiones Multidimensionales constituyen un lenguaje de consulta para Base de Datos Multidimensionales sobre cubos OLAP y son utilizadas en Inteligencia de Negocios para generar reportes que facilitan la toma de decisiones basados en datos históricos, con la posibilidad de cambiar la estructura o rotar el cubo. Una consulta MDX es muy similar a una consulta SQL, ya que devuelve un conjunto de celdas, que es un subconjunto de celdas del cubo original.

Para cada reporte se define una consulta MDX, por lo que se tienen 13 consultas MDX, ya que este el número de reportes candidatos. En el **Anexo 15** se relacionan las consultas MDX por reportes, pertenecientes al libro de trabajo Asistencia Social.

Al ejecutar la consulta MDX en el Pentaho BI Server, que corresponde al reporte **TS1 – Asistencia Social. Indicadores Seleccionados en Cuba**, se visualiza el contenido del mismo.



	Años	
Indicadores	2006	2007
Gastos por la asistencia social	572.386	590.7
Beneficiarios de la asistencia social	599,505	595,181
Núcleos protegidos por la asistencia social	328,462	334,692
Adultos mayores beneficiarios de la asistencia social	143,483	145,275
Personas con discapacidad beneficiaria de la asistencia social	97,347	98,727
Madres de hijos con discapacidad severa beneficiarias de la asistencia social	7,462	7,789
Beneficiarios del servicio de asistente social a domicilio	13,537	16,182

Slicer: [Measure=valor_real]

Ilustración 18: Reporte TS1 - Asistencia Social. Indicadores Seleccionados en Cuba.

3.3.2 Configurar la seguridad de los usuarios

Otro aspecto de vital importancia para completar el subsistema de visualización es configurar la seguridad de los usuarios en el sistema. Esto implica la creación de usuarios, roles y la asignación de permisos a los usuarios y roles creados.

Para el desarrollo del subsistema de visualización del Mercado de datos “Protección del Trabajo y Seguridad Social” se crean y configuran dos roles y dos usuarios que poseen diferentes niveles de acceso al sistema.

El rol de administrador posee todos los permisos sobre el sistema y tiene asignado el usuario admin.

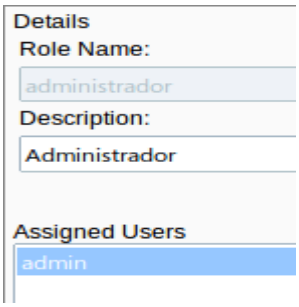


Ilustración 19: Rol y Usuario de Administración.

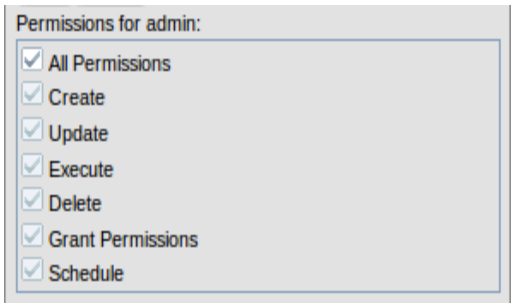


Ilustración 20: Permisos asignados al usuario “admin”.

El rol de Análisis posee única y exclusivamente permisos de ejecución sobre el sistema y tiene asignado el usuario analista.

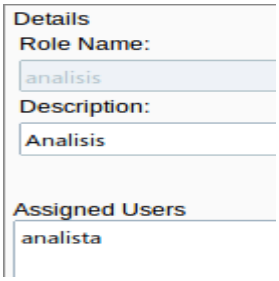


Ilustración 21: Rol y Usuario de Análisis.

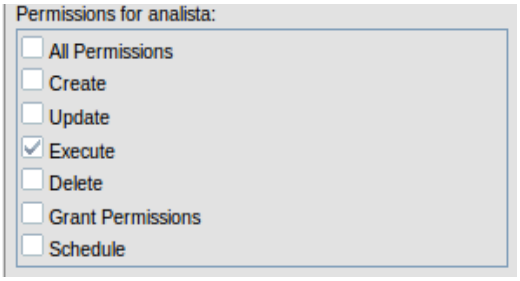


Ilustración 22: Permisos asignados al usuario "analista".

Conclusiones del capítulo

Una vez finalizado este capítulo y con ello la implementación del Mercado de datos "Protección del Trabajo y Seguridad Social" se puede concluir que:

- La implementación del trabajo y las cuatro transformaciones para cargar los datos de los ficheros fuentes garantizan la consistencia y fiabilidad de los datos.
- Los reportes candidatos implementados garantizan la disponibilidad de la información de manera eficiente y organizada, facilitando el proceso de toma de decisiones.
- La configuración de la seguridad, tanto a nivel de Base de Datos como a nivel de aplicación, garantiza la disponibilidad, confiabilidad e integridad de la información.

Capítulo 4. Validación y pruebas al Mercado de datos

Introducción

El Mercado de datos al entrar en contacto con los usuarios finales, entra en un ciclo iterativo e incremental, de lo simple a lo complejo, donde el sistema nunca descansará puesto que a él son adheridos, con el transcurso del tiempo, nuevos años de información, procesos de negocios de la empresa, nuevas necesidades o insatisfacciones del cliente. En el momento en que se implanta en la empresa y al entrar en plena explotación, el Mercado de datos crece ilimitadamente, al ser alimentado con los datos históricos, a la vez que se vuelve complejo, y es cuando comienzan a observarse los beneficios de los tiempos de respuesta, el dinamismo en la elaboración de los reportes, los conocimientos que puedan ser extraídos de la información almacenada y la efectiva preparación de los usuarios finales, garantizan así el éxito del Sistema.

4.1 Pruebas

Las pruebas son una actividad en la cual un sistema o componente es ejecutado bajo unas condiciones o requerimiento específicos y los resultados son observados y registrados. Esta actividad no garantiza la ausencia de defecto; solo puede demostrar que existen, permitiendo evaluar la calidad de software.

Para evaluar la calidad del Mercado de datos “Protección del Trabajo y Seguridad Social” se utilizó el **Modelo V** definido por DATEC, con el fin de lograr que el producto cumpla con las especificaciones del negocio y tenga una mayor calidad.

El modelo en V muestra cómo se relacionan las actividades de prueba con el análisis y el diseño. Ver **Ilustración 24**.

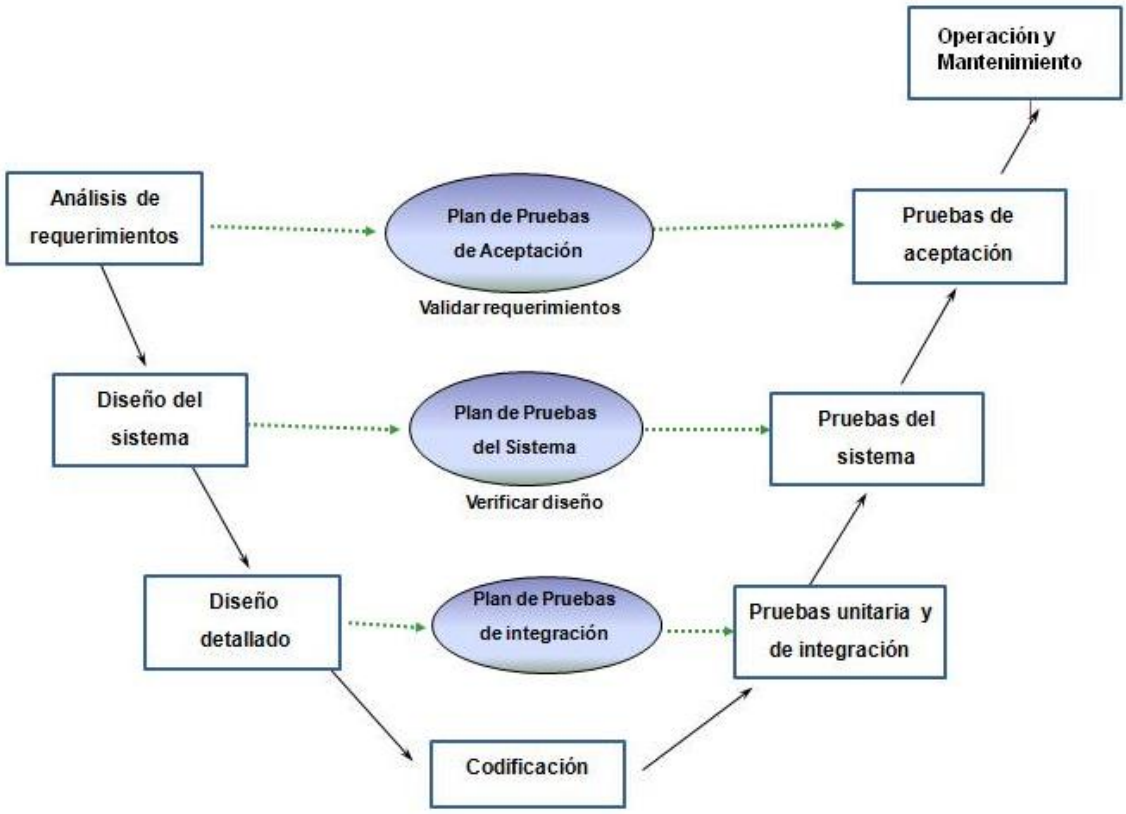


Ilustración 23: Modelo V.

Para realizar las pruebas al Mercado de datos se confeccionó una lista de chequeo y se diseñaron cuatro casos de pruebas.

4.1.1 Lista de Chequeo

La lista de chequeo es un listado de preguntas, en forma de cuestionario que se utiliza para verificar el grado de cumplimiento de determinas reglas establecidas con un fin determinado. Esta tiene como objetivo evaluar la eficiencia del Mercado de datos, donde se mide una serie de indicadores que se encuentran implicados en el proceso de la creación de la capa de integración y visualización del sistema, además de medir, la calidad de los artefactos y documentos generados en la realización del producto.

En esta todos los indicadores que contiene se encuentran distribuidos en tres secciones fundamentales:

- Estructura del documento: abarca todos los aspectos definidos por el expediente de proyecto o el formato establecido por el proyecto.

- Indicadores definidos: abarca todos los indicadores a evaluar durante la etapa de desarrollo del mercado.
- Semántica del documento: contempla todos los indicadores a evaluar respecto a la ortografía, redacción y demás.
- Los elementos que forman parte de la estructura de la lista de chequeo son:
- Peso: define si el indicador a evaluar es crítico o no. El mismo se define con una C si es crítico.
- Indicadores a evaluar: son los indicadores a evaluar en las secciones Estructura del documento, Semántica del documento e Indicadores definidos por la etapa
- Evaluación (Eval): es la forma de evaluar el indicador en cuestión. El mismo se evalúa de uno en caso de que exista alguna dificultad sobre el indicador y en caso de que el indicador revisado no presente problemas.
- N.P. (No Procede): se usa para especificar que el indicador no es necesario evaluarlo en ese caso.
- Cantidad de elementos afectados (CEA): especifica la cantidad de errores encontrados sobre el mismo indicador.
- Comentario (Comt): especifica los señalamientos o sugerencias que quiera incluir la persona que aplica la lista de chequeo. Pueden o no existir señalamientos o sugerencias.

Una vez definida la estructura de la lista de chequeo (ver **Anexo 16**), se aplica al Mercado de datos y se genera el siguiente gráfico de barras (ver **Ilustración 25**), donde se visualiza el comportamiento de los 25 indicadores, de los cuales 11 son críticos.

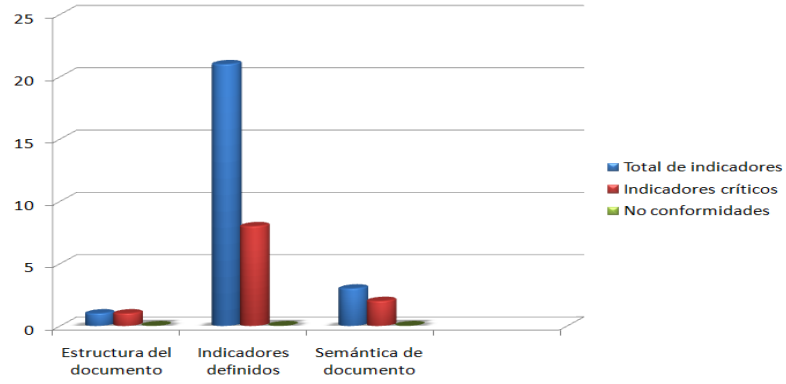


Ilustración 24: Comportamiento de los Indicadores de la Lista de Chequeo.

4.1.2 Casos de Pruebas

Los casos de prueba permiten verificar la calidad del software, estos son utilizados para identificar posibles fallos de implementación y comprobar el grado de cumplimiento de las especificaciones iniciales del sistema.

En el Mercado de datos “Protección del Trabajo y Seguridad Social” se aplicaron cuatro casos de prueba, uno a cada caso de uso de información identificados en la etapa de análisis. En el **Anexo 17** se relaciona el caso de prueba correspondiente al caso de uso “Analizar indicadores de Pensionados”.

Los casos de pruebas se están constituidos por la sección **Reportes candidatos**, donde se verifica que los reportes se visualicen con las variables correspondientes.

4.1.3 Pruebas de Volumen y Carga

Las pruebas que poseen relación con el rendimiento, capacidad y concurrencia son las que más impactan en el desarrollo de los Almacenes de datos, razón por la cual, las pruebas de volumen y carga también serán aplicadas al Mercado de datos.

4.1.3.1 Pruebas de volumen

Las pruebas de volumen son pruebas típicas de entornos que utilizan Bases de Datos. Las mismas se realizan para analizar el comportamiento del sistema o Base de Datos con volúmenes de datos almacenados lo más similar posible a los esperados en la explotación real del sistema. (7) Para el sistema en cuestión la Base de Datos se pobló con los datos reales suministrados por los especialistas de la ONE, lo que implica que el tamaño es muy similar al real esperado.

Al introducir los datos no se presentaron problemas de límite de capacidad, ni de volumen de datos. Tampoco se detectaron desbordamientos de matrices, 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 gestor utilizado y el diseño de las estructuras de la Base de Datos implementadas soportan completamente el almacenamiento de los niveles de información requeridos para la puesta en producción del Mercado de datos.

4.1.3.2 Pruebas de Carga

Las pruebas de carga consisten en someter a una aplicación y/o Base de Datos a un régimen de carga de trabajo (habitualmente por simulación de concurrencia) similar al esperado en la explotación real del sistema. El objetivo de estas pruebas es buscar consultas mal diseñadas, consultas candidatas a

optimización, la necesidad de índices adicionales, código mal diseñado, tiempo de demora de respuesta de magnitudes inaceptables, hardware insuficiente, problemas de control de concurrencia, etc. (7)

Para realizar las pruebas se utiliza la herramienta Apache-Jakarta JMeter por ser un generador de carga diseñado para la realización de pruebas de carga y stress, corre sobre la máquina virtual de java por lo que es multiplataforma. Genera carga por diversos protocolos: FTP (Protocolo de Transferencia de Archivos), HTTP (Protocolo para la Transferencia de Hipertextos), HTTPS (HTTP Seguro), SQL, etc. Además maneja cookies y autenticación, realiza carga variable, en niveles de concurrencia, número de veces, tiempo, etc.; y su característica principal radica en que pertenece a la familia de software libre. (7)

La arquitectura general que se utiliza para la realización de las pruebas serán tres estaciones clientes con el JMeter configurado directamente con el servidor de Base de Datos (ver **Ilustración 26**). Dos de las estaciones clientes sólo limitarán su uso a realizar peticiones indefinidamente al servidor y la otra para llevar las estadísticas con el número de muestras definido en 30. Se le realizan pruebas con 10 y 20 usuarios concurrentes debido a que según los especialistas de la ONE, nunca existirán más de 10 usuarios registrados en el servidor, por lo que la concurrencia será mínima. Además, las consultas se realizarán sobre la tabla de hechos.

Ilustración 25: Configuración de las Pruebas de Carga.

Prueba No 1: Comportamiento de los indicadores de los Pensionados según la medida valor real.

Dimensiones involucradas: División Política Administrativa, Temporal Año e Indicador General.

Cantidad total de tuplas en la tabla de hechos: 30 tuplas

Cantidad total de tuplas recuperadas: 30 tuplas

Consulta: SELECT * FROM mart_trab_seg_social.hech_pensionados WHERE valor_real >= 0;

Cantidad de usuarios	Media	Mediana	Mínimo	Máximo
10 usuarios concurrentes	68	68	63	78
20 usuarios concurrentes	90	90	70	124

Tabla 9: Informe de las pruebas de los Pensionados.

Prueba No 2: Comportamiento de los indicadores de la Asistencia Social según la medida valor_real.

Dimensiones involucradas: Provincia, Temporal Año e Indicador General.

Cantidad total de tuplas en la tabla de hechos: 210 tuplas

Cantidad total de tuplas recuperadas: 210 tuplas

Consulta: SELECT * FROM mart_trab_seg_social.hech_asistencia_social WHERE valor_real >= 0;

Cantidad de usuarios	Media	Mediana	Mínimo	Máximo
10 usuarios concurrentes	51	53	15	73
20 usuarios concurrentes	121	143	25	165

Tabla 10: Informe de las pruebas de la Asistencia Social.

Prueba No 3: Comportamiento de los indicadores de la Seguridad Social según la medida valor_actual.

Dimensiones involucradas: División Política Administrativa, Temporal Semestre, Nomenclador de Actividad Económica, Clasificador de Actividad Económica e Indicador General.

Cantidad total de tuplas en la tabla de hechos: 59 551 tuplas

Cantidad total de tuplas recuperadas: 59 551 tuplas

Consulta: SELECT * FROM mart_trab_seg_social.hech_seguridad_social WHERE valor_actual>= 0;

Cantidad de usuarios	Media	Mediana	Mínimo	Máximo
10 usuarios concurrentes	11 131	11 951	5 501	11 989
20 usuarios concurrentes	85 958	92 202	42 356	93 906

Tabla 11: Informe de las pruebas de la Seguridad Social.

Prueba No 4: Comportamiento de los indicadores de la Protección del Trabajo según la medida valor_actual.

Dimensiones involucradas: División Política Administrativa, Temporal Trimestre, Nomenclador de Actividad Económica, Clasificador de Actividad Económica e Indicador General.

Cantidad total de tuplas en la tabla de hechos: 14 278 tuplas

Cantidad total de tuplas recuperadas: 14 278 tuplas

Consulta: SELECT * FROM mart_trab_seg_social.hech_proteccion_trabajo WHERE valor_actual >= 0;

Cantidad de usuarios	Media	Mediana	Mínimo	Máximo
10 usuarios concurrentes	887	945	401	1 021
20 usuarios concurrentes	1 187	1 373	704	1 558

Tabla 12: Informe de las pruebas de la Protección del Trabajo.

Como puede apreciarse los resultados se enmarcan en cuatro variables, aportadas por la herramienta JMeter (ver **Anexo 18**) de significativo valor para la presentación de los resultados de las pruebas: la **media**, valor de la suma aritmética de los tiempos de respuesta dividido entre dos, la **mediana**, valor de la variable que deja el mismo número de datos antes y después que él, una vez ordenados estos, de acuerdo con esta definición el conjunto de datos menores o iguales que la mediana representarán el 50% de los datos, y los que sean mayores que la mediana representarán el otro 50% del total de datos de la muestra; el valor **mínimo**, que se refiere al tiempo de respuesta menor de todos los usuarios que hicieron peticiones concurrentes y el valor **máximo** que, similarmente, es el tiempo mayor de respuesta a todos los usuarios. (7)

En las tablas anteriores los tiempos de respuestas, dados en milisegundos, oscilaron en dependencia de la cantidad de filas que se recuperen en la consulta. En general los resultados obtenidos son satisfactorios evidenciándose un aumento casi simétrico entre cada una de las configuraciones realizadas, destacándose los mayores tiempos en la tabla de hechos hech_seguridad_social debido a que la consulta devuelve más cantidad de tuplas a cada usuario, convirtiéndose en un proceso un poco más lento.

4.2 Validación del Sistema

Después de haber concluido el ciclo de desarrollo del Mercado de datos “Protección del Trabajo y Seguridad Social”, en el cual quedan definidos aspectos importantes referentes al análisis, diseño e implementación del mismo, corresponde evaluar y validar el Sistema, donde la participación de los clientes es de suma importancia en la etapa.

Resulta de gran importancia que los clientes estén inmersos en esta etapa de evaluación y validación del sistema, pues:

- Pueden ser encontradas discrepancias con los requerimientos identificados en la etapa de análisis.
- La familiarización con el ambiente de explotación de la información.
- Para refinar el Sistema en función de que quede lo más completo posible.

En el proyecto esta etapa duró aproximadamente dos meses donde estuvo involucrada la principal cliente del mismo que es la Jefa de Informática de la ONE Ing. Elena Fernández García mediante una sistema de chequeo semanal donde se le presentaba los principales resultados y se definían las posibles funcionalidades a agregar.

Uno de los puntos fundamentales en este proceso fueron los tipos de reportes que la ONE realiza en su quehacer diario, por lo que los especialistas se encargaron de recopilar un variado y amplio conjunto de tablas de salida, que ayudaron de sobremanera para enfocar el diseño de las estructuras hacia los principales pedidos de información.

Para el logro del éxito en esta etapa ha tenido un valor significativo la disposición, colaboración y asistencia de los especialistas de la ONE que en todo momento han brindado la ayuda necesaria para la resolución de los problemas que fueron apareciendo a medida que se refinaba el Sistema.

El hecho de que el Sistema haya cumplido los objetivos propuestos inicialmente y satisfecho total o mayoritariamente los requisitos definidos, no significa que ya esté apto para ser montado e instalado en la entidad, sino que es importante también la preparación de los especialistas que interactuarán con el mismo.

Finalmente se evidencia que el proceso de validación del Sistema se ha realizado satisfactoriamente, hecho que ha quedado plasmado en la carta de aceptación por parte de cliente. Por tal motivo se puede decir que los objetivos propuestos han sido cumplidos y las expectativas que se tenían con el Mercado de datos han sido superadas.

Conclusiones del capítulo

Después de analizar los resultados obtenidos en la etapa de validación se concluye que:

- Las pruebas de volumen validaron la infraestructura de hardware y software propuestas garantizando la capacidad de gestión de los datos almacenados.
- Las pruebas de carga resultaron una herramienta eficaz en el proceso de optimización y demostraron que la estrategia de indexado cumple con los tiempos de respuestas aceptables para este tipo de solución.

Conclusiones

La investigación cumplió los objetivos planteados y se arribaron a las siguientes conclusiones:

- El análisis y diseño del Mercado de datos cumplió con todo lo requerido en el negocio, satisfaciendo las necesidades del cliente y posibilitando una implementación robusta del sistema.
- Se modelaron e implementaron las estructuras dimensionales requeridas para el manejo de la información relevante del modelo estadístico M5201, soportando el proceso de toma de decisiones.
- La implementación de los tres subsistemas condujo a la obtención de un Mercado de datos poblado y funcional, con la información disponible para ser consultada por parte de los usuarios, apoyando el proceso de toma de decisiones.
- Las pruebas realizadas permitieron validar la calidad de la solución propuesta, obteniendo resultados satisfactorios en cada una de ellas.

Recomendaciones

Con el propósito de enriquecer la propuesta realizada en este trabajo, se sugiere:

- Optimizar las consultas de recuperación de información en el Sistema Gestor de Bases de Datos.
- Migrar a la versión superior de PostgreSQL 9.x con vista a utilizar las nuevas funcionalidades que esta ofrece.
- Mejorar las políticas de seguridad implementadas.
- Implementar técnicas de Minería de datos que posibiliten mejorar el proceso de toma de decisiones.

Referencias bibliográficas

1. **Ralph Kimball y Margy Ross.** *The Data Warehouse Toolkit.* E.U.A : Wiley Publishing Inc, 2002.

2. **Inmon, William H.** *Building the Data Warehouse.* E.U.A : Wiley Publishing Inc, 2005.

3. **Claudia Imhoff, Nicholas Galemmo y Jonathan G. Geiger.** *Mastering Data Warehouse Desing, Relational and Dimentional Techniques.* E.U.A : Wiley Publishing Inc, 2003.

4. **Sinnexus.** Data Warehouse. [En línea] [Citado el: 15 de Noviembre de 2010.] http://www.sinnexus.com/business_intelligence/datawarehouse.aspx.

5. **Peñaloza, Lucía Victoria Hernández.** *Tesis para logra el título de Magíster: Diseño y Construcción de un Data Mart para la mantención de Indicadores de Sostenibilidad de la Industria del Salmón.* Chile : s.n, 2008.

6. **Ponniah, Paulraj.** *Data Warehousing Fundamentals.* E.U.A : Wiley Publishing Inc, 2001.

7. **Sierra, Julio Ernesto Ortiz.** *Tesis: Diseño e Implementación de un Mercado de Datos para la Oficina Nacional de Estadísticas.* Cuba : s.n, 2009.

8. **Lilian Hobbs y otros.** *Oracle Database 10g Data Warehousing.* E.U.A : ELSEVIER Digital Press, 2005.

9. **Ralph Kimball y otros.** *The Data Warehouse Lifecycle Toolkit.* E.U.A : Wiley Publishing Inc.

10. **Cordero, Vladimir Urquia.** *Tesis: Arquitectura y Componente de Almacenamiento del ODS para SIIPOL.* Cuba : s.n, 2010.

11. **Oracle.** Sitio Oficial de Oracle. [En línea] http://www.oracle.com/solutions/business_intelligence/feature_dw_leadership.html.

12. **NAKAMURA, Y.** *Sistemas Gestores de Bases de Datos. Revista de Posgrado.* México : Universidad Autónoma de México, 2007.

13. **Inocencio, I. B.** Nueva Economía, Internet y tecnología. [En línea] Julio de 2004. [Citado el: 10 de Noviembre de 2010.] http://www.gestiopolis.com/canales2/gerencia/1/busint.htm.

14. **Keyrus.** Arquitectura de una Solución de BI. [En línea] [Citado el: 15 de Diciembre de 2010.] http://www.keyrus.com.

15. **Curto, Josep.** *CIF vs MD Dos enfoques clásicos en el diseño de la arquitectura de un Data Warehouse.* España : s.n, 2008.

16. **Garavito, Julio.** *Manual Básico de PostgreSQL.* Colombia : Escuela Colombiana de Ingeniería, 2007.

Bibliografía

1. **Ralph Kimball y Margy Ross.** *The Data Warehouse Toolkit.* E.U.A : Wiley Publishing Inc, 2002.

2. **Inmon, William H.** *Building the Data Warehouse.* E.U.A : Wiley Publishing Inc, 2005.

3. **Claudia Imhoff, Nicholas Galemmo y Jonathan G. Geiger.** *Mastering Data Warehouse Desing, Relational and Dimentional Techniques.* E.U.A : Wiley Publishing Inc, 2003.

4. **Sinnexus.** Data Warehouse. [En línea] [Citado el: 15 de Noviembre de 2010.] http://www.sinnexus.com/business_intelligence/datawarehouse.aspx.

5. **Peñaloza, Lucía Victoria Hernández.** *Tesis para logra el título de Magíster: Diseño y Construcción de un Data Mart para la mantención de Indicadores de Sostenibilidad de la Industria del Salmón.* Chile : s.n, 2008.

6. **Ponniah, Paulraj.** *Data Warehousing Fundamentals.* E.U.A : Wiley Publishing Inc, 2001.

7. **Sierra, Julio Ernesto Ortiz.** *Tesis: Diseño e Implementación de un Mercado de Datos para la Oficina Nacional de Estadísticas.* Cuba : s.n, 2009.

8. **Lilian Hobbs y otros.** *Oracle Database 10g Data Warehousing.* E.U.A : ELSEVIER Digital Press, 2005.

9. **Ralph Kimball y otros.** *The Data Warehouse Lifecycle Toolkit.* E.U.A : Wiley Publishing Inc.

10. **Cordero, Vladimir Urquia.** *Tesis: Arquitectura y Componente de Almacenamiento del ODS para SIIPOL.* Cuba : s.n, 2010.

11. **Oracle.** Sitio Oficial de Oracle. [En línea] http://www.oracle.com/solutions/business_intelligence/feature_dw_leadership.html.

12. **NAKAMURA, Y.** *Sistemas Gestores de Bases de Datos. Revista de Posgrado.* México : Universidad Autónoma de México, 2007.

13. **Inocencio, I. B.** Nueva Economía, Internet y tecnología. [En línea] Julio de 2004. [Citado el: 10 de Noviembre de 2010.] http://www.gestiopolis.com/canales2/gerencia/1/busint.htm.

14. **Keyrus.** Arquitectura de una Solución de BI. [En línea] [Citado el: 15 de Diciembre de 2010.] http://www.keyrus.com.

15. **Curto, Josep.** *CIF vs MD Dos enfoques clásicos en el diseño de la arquitectura de un Data Warehouse.* España : s.n, 2008.

16. **Garavito, Julio.** *Manual Básico de PostgreSQL.* Colombia : Escuela Colombiana de Ingeniería, 2007.

17. **Zenaido, Rosendo.** *Borrador Tesis de Doctorado: Metodología para el Diseño de Almacenes de Datos.* España : s.n, 2008.

18. **Wang, John.** *Encyclopedia of Warehousing and Mining.* E.U.A : Idea Group Reference, 2006.

19. **Velasco, Roberto Hernando.** rhernando.net. [En línea] [Citado el: 15 de dicieembre de 2010.] htttp://www.rhernando.net/modules/tutorials/doc/bd/dw.html.

20. **Lockhart, Thomas.** *Tutorial de PostgreSQL.* s.l. : s.n, 2000.

21. **Iznaga, Yonelbys.** *Tesis: Sistema Data Warehouse.* Cuba : s.n, 2008.

22. **Huamantumba, Rayner.** *Manual para diseño y desarrollo de Data Mart.* s.l. : s.n, 2007.

23. **Chuc-Durán, Diana Graciela.** *Introducción a los Datawarehouses.* México : s.n, 2007.

24. **Bolaños, Carlos.** *Base de Datos para redes.* Universidad Don Bosco : Facultad de Estudios Tecnológicos, 2009.

25. **Bernabeu, Ricardo Dario.** *Hefesto: Metodología propia para la construcción de un Data Warehouse.* Argentina : s.n, 2007.

26. **Adamson, Christopher.** *Mastering Data Warehouse Aggregates.* E.U.A : Wiley Publishing Inc, 2006.

27. **Ken England y Gavy Powell.** *Performance, Optimization and Tunning handbook.* E.U.A : ElSEVIER Inc, 2007.

28. **William H. Inmon, Strauss y Neushloss.** *DW 2.0 The Architecture for de Next Generation.* E.U.A : Wiley Publishing Inc,, 2007.

29. **Poole y otros.** *Common Warehouse Metamodel.* E.U.A : Wiley Publishing Inc, 2003.

30. **Hurtado y otros.** *Base de Datos y Data Warehouse: Herramientas Estratégicas para la eficacia comercial.* Granada : s.n, 2003.