Universidad de las Ciencias Informáticas
Facultad 6

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

Título: "Mercado de datos Series históricas de industria para el Sistema de Información de Gobierno."

Autor:
Eddy David Amaya Abreu

Tutores:
Ing. José Leandro González González.
Ing. Arodys Eugenio Dominguez Vaillant.

La Habana, Junio de 2012
"Año 54 de la Revolución"

Muchos me dirán aventurero, y lo soy, solo que de un tipo diferente y de los que ponen el pellejo para demostrar sus verdades…

Ché.

DECLARACION DE AUTORÍA

Declaro ser autor de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

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

_____

Eddy David Amaya Abreu

Autor

_____ _____

Ing. José Leandro González González.

Tutor

Ing. Arodys Eugenio Dominguez Vaillant.

Tutor

DATOS DE CONTACTO

Ing. José Leandro González González:

Graduado en Ingeniería en Ciencias Informáticas en el 2008 en la Universidad de las Ciencias Informáticas (UCI). Actualmente trabaja en el Centro de Tecnología de Gestión de Datos (DATEC) adjunto al Departamento de Sistemas Digitales, Facultad 6.

Ing. Arodys Eugenio Dominguez Vaillant:

Graduado en Ingeniería en Ciencias Informáticas en el 2007 en la Universidad de las Ciencias Informáticas (UCI). Actualmente trabaja en el Departamento de Ciencias Básicas adjunto al Centro de Tecnología de Gestión de Datos (DATEC), Facultad 6.

Agradecimientos

** *A mis padres por estar a mi lado en cada momento apoyándome incondicionalmente, por su papel significativo en el logro de mis metas y por ser la mayor inspiración.*

** *A mis abuelos que están y a los no están físicamente hoy conmigo, por favorecer mi cuidado y formación.*

** *A mis hermanas por la gran responsabilidad que tengo de ser cada día mejor para guiarlas y vean en mí un gran ejemplo a seguir.*

** *A mi futura esposa por formar parte de mi vida, por contribuir en la realización de este sueño, por su amor, su cariño, su comprensión, su paciencia y su apoyo constante.*

** *A mi familia por su apoyo durante toda mi carrera.*

** *A mis tutores por sus persistentes aportes en la finalización de este trabajo.*

** *A mis compañeros y al claustro de profesores que han colaborado con mi formación pre-profesional.*

** *A todas las personas que me han demostrado su amistad, que me han ayudado cuando más lo he necesitado y que han contribuido a ser este sueño realidad.*

*** *En fin a todos muchas gracias...* ***

Dedicatoria

A mis padres por guiarme en la vida, por estar con los brazos abiertos siempre que los necesito, por todo el cariño que me regalan, por todos sus esfuerzos y sacrificios para lograr mi desarrollo profesional. A mis hermanas que son las personitas que me dan la inspiración para ser mejor cada día, a mi ahijado para que un día se sienta orgulloso de mí y a mi novia que me ha apoyado en todos los momentos difíciles.

A todos, hagan suyo este sueño hecho realidad…

Resumen

Con el avance de las Tecnologías de la Información y las Comunicaciones ha aumentado el nivel de procesamiento y almacenamiento de la información, existiendo diversos mecanismos a la hora de almacenar y analizar la información. Surgen así los mercados de datos, tecnología en la que está enmarcada la investigación. Este trabajo surge a través de la colaboración existente entre el Centro de Tecnologías y Gestión de Datos y la Oficina Nacional de Estadísticas e Información, a partir de la necesidad de analizar el comportamiento de los indicadores del área Industria para apoyar a la toma de decisiones respecto a los procesos industriales. Teniendo como objetivo desarrollar un mercado de datos para el sistema de información de gobierno en el área Industria en aras de obtener un mejor desempeño en esta área. Se obtuvo una solución integrada donde se incluyen tres subsistemas: almacenamiento, integración y visualización; lo que permite realizar un mejor análisis de la información a través de los reportes, cruces de variables y demás aspectos de interés tratados en el área antes especificada.

Palabras Claves:

Tecnologías de la Información y las Comunicaciones (TIC)
Oficina Nacional de Estadística e Información (ONEI)
Centro de Tecnologías y Gestión de Datos (DATEC)
Almacenes de datos (AD)
Mercado de datos (MD)

Índice de contenidos

Índice de figuras

Índice de tablas

Introducción

Hoy día se puede encontrar mucha información asociada a diversos datos, que en muchos casos no ofrecen información a simple vista, la diversidad de fuentes de información y la variabilidad de los formatos de la misma ha conllevado a pensar que dicha información pueda ser centralizada. En el mundo se conocen diversas herramientas realizadas a la medida y otras de carácter genérico que son capaces de favorecer la toma de decisiones de una empresa o entidad a la cual se le incorporan. La inteligencia de negocio ha sido un tema muy difundido dentro de los futuros empresariales a nivel mundial. En la actualidad, la mayor parte de las encuestas de la industria muestran que la mayoría de los usuarios de Almacenes de datos (AD) se apoyan en hojas de cálculo, herramientas para elaboración de informes y análisis, o en sus propias aplicaciones personalizadas para recuperar datos de los almacenes y formatearlos a informes y gráficas para la empresa. Durante casi dos décadas, las bases de datos multidimensionales y sus sistemas de exposición de información analítica han proporcionado presentaciones de ventas y demostraciones atractivas en las ferias comerciales. Los servidores y las herramientas de escritorio de Procesamiento analítico en línea (OLAP) ofrecen soporte al análisis de alta velocidad de datos con relaciones complejas, como por ejemplo combinaciones de los productos de una industria. En Cuba hace algunos años se vienen realizando investigaciones relacionadas con el tema y precisamente una de estas tiene lugar en la Oficina Nacional de Estadística e Información (ONEI).

La ONEI, órgano rector en cuanto a estadística e información representa en Cuba, tiene como objetivo principal conseguir datos estadísticos con la mayor calidad posible innatos del Sistema de Información Estadístico Nacional (SIEN). Dicho órgano trabaja con innumerables modelos estadísticos en los que se recoge la información referente a los distintos sectores de la economía y la sociedad, dentro de los cuales se encuentra la Industria. Este en coordinación con dicho sector almacena gran cantidad de información estadística año tras año, en ella aparecen las Series históricas. Aclarar que dicho sector al que se refiere dentro de este organismo se tratará como área de la entidad antes mencionada. La información correspondiente se almacena en diferentes modelos por las entidades existentes a lo largo del territorio nacional, posteriormente son enviados a las entidades municipales de la ONEI correspondientes a cada territorio y luego se envían a las sedes provinciales; por último, la información es digitalizada en ficheros con formato ¨dbf¨ y remitidos a la oficina central que se sitúa en Paseo No. 60, e/ 3ra y 5ta, Vedado, Plaza de la Revolución, La Habana.

La Universidad de la Ciencias Informáticas (UCI) es una institución de educación superior que está inmersa en el proceso docente-productivo, siendo esta un centro de producción de software. La Universidad la constituyen un grupo de facultades que a su vez contienen centros de desarrollo, uno

de ellos es el Centro de Tecnología de Gestión de Datos (DATEC). En el centro existe una colaboración con la Oficina Nacional de Estadística e Información (ONEI) para contribuir el proceso de la toma de decisiones, a través de la misma se creó el Sistema de Información de Gobierno (SIGOB).

Actualmente no se han encontrado aplicaciones encargadas de gestionar la información de Industria, en cuanto a este tema en específico.

En la actualidad existen diversos problemas en el trabajo con las Series históricas en el área de Industria, dichos problemas vienen aparejados a la compleja manipulación de los datos, debido a que se almacenan en formato Excel. Se generan ficheros anuales generando dificultades en la obtención de información estadística a partir de los datos contenidos en dichos archivos. La existencia de múltiples versiones de los datos se interpone a la posibilidad de que estos se encuentren integrados y que posean buena calidad, lo que facilita que puedan ser analizados. Además en dicha oficina no existe una aplicación informática que permita generar reportes relacionados con las Series históricas de industria, a través de los cuales se pueda realizar un análisis de la información, por lo que los procesos de recuperación y elaboración de informes son más costosos en esfuerzo y tiempo dedicado.

En la búsqueda de mejoras en las formas de almacenar, recuperar y presentar la información proveniente de los organismos, específicamente como reportes principales, cruces de variables, indicadores, porcentajes y demás aspectos de interés constituye una necesidad urgente para aumentar la disponibilidad de información y el proceso de toma de decisiones del área antes mencionada.

Teniendo en cuenta lo anteriormente expuesto se plantea el siguiente **problema de la investigación**: ¿Cómo contribuir al proceso de toma de decisiones en el área de las Series históricas de industria para el Sistema de Información de Gobierno?

Se define como **objeto de estudio** los Almacenes de datos y se delimita como **campo de acción** mercado de datos Series históricas de industria para el Sistema de Información de Gobierno.

El **objetivo general** de la presente investigación es: Desarrollar el mercado de datos Series históricas de industria para el Sistema de Información de Gobierno, que contribuya a la toma de decisiones.

A partir del objetivo general se desglosan los siguientes **objetivos específicos**:

- Fundamentar la selección de la metodología, herramientas y tecnologías a utilizar en el desarrollo del mercado de datos Series históricas de industria.
- Realizar el análisis y diseño del mercado de datos Series históricas de industria.
- Implementar el mercado de datos Series históricas de industria.
- Realizar pruebas al mercado de datos Series históricas de industria.

Para el cumplimiento a los mismos se plantean las **tareas de investigación** que a continuación se muestran:

- Caracterización de la metodología, herramientas y tecnologías a utilizar en el desarrollo de AD.

- Levantamiento de requisitos.
- Descripción de los casos de uso del MD.
- Definición de los hechos, las medidas y las dimensiones del MD.
- Diseño del modelo de datos.
- Implementación del subsistema de almacenamiento.
- Definición de la arquitectura del MD.
- Diseño del subsistema de integración.
- Diseño del subsistema de visualización.
- Diseño de los casos de prueba.
- Implementación del subsistema de integración.
- Implementación del subsistema de visualización.
- Aplicación de las listas de chequeo.
- Aplicación de los casos de prueba.

Estructura de la Tesis

El presente trabajo de diploma estará estructurado de la siguiente forma: introducción, cuatro capítulos, conclusiones, recomendaciones, referencias bibliográficas, bibliografía, anexos y glosario de términos.

Capítulo 1: Fundamentación Teórica de los Almacenes de datos.

En el presente capítulo se profundizará en la fundamentación teórica de la investigación. Además, se realizará un estudio de las herramientas y metodología de desarrollo utilizada, así como de los procesos de integración de los datos e inteligencia de negocio que forman parte del desarrollo de un Almacén de datos (AD).

Capítulo 2: Análisis y diseño del mercado de datos.

En este capítulo se realizará un estudio preliminar del negocio, la necesidades de información, las reglas del negocio, descripción de los casos de uso, identificación de dimensiones, hechos y medidas, desarrollo de la matriz BUS, modelo de datos, arquitectura de información, así como definir los reportes candidatos y el diseño de los procesos de integración.

Capítulo 3: Implementación del mercado de datos.

En el mismo se implementará el proceso de extracción, transformación y carga (ETL), y la capa de inteligencia del negocio para el área de la Industria en la ONEI, específicamente las Series históricas de dicha área, teniendo en cuenta las necesidades del cliente.

Capítulo 4: Validación del mercado de datos.

En el presente capítulo se realizará la validación de la solución propuesta a través de las distintas vías, para así verificar su correcto funcionamiento y que cuente con la calidad requerida por el cliente. Esta validación se realiza a partir de la aplicación de listas de chequeo y los casos de prueba diseñados, aplicados a la aplicación.

Capítulo 1: Fundamentación Teórica de los Almacenes de datos

Introducción

En el presente capítulo se profundizará en la fundamentación teórica de la investigación. Además, se realizará un estudio de las herramientas y metodología de desarrollo utilizada, así como de los procesos de integración de los datos e inteligencia de negocio que forman parte del desarrollo de un AD.

1.1 Industria

Los primeros esfuerzos en la economía se dirigieron objetivamente hacia el logro de la industrialización de la nación, dando a las fuerzas productivas la oportunidad de dilatar su potencial. Los pasos tentativos se iniciaron con la constitución del Ministerio de Industrias, donde se concentraron inicialmente todas las instancias del sector industrial, convertido en una sola entidad y bajo la misma dirección.

Este sector se encuentra identificado en la ONEI con el Departamento de Industria donde se recopila información brindada por tal ministerio como organismo rector, siendo emitida por todas las entidades territoriales. Dicho departamento cuenta con modelos estadísticos donde se recoge gran cantidad de datos.

1.2 Aplicaciones existentes para el trabajo con las Series históricas

La ONEI y el SIEN en coordinación con el SIGOB desarrollan AD que contribuyen a la toma de decisiones, donde su principal objetivo es almacenar toda la información referente de los organismos territoriales a lo largo de todo el país. Actualmente no se han encontrado aplicaciones encargadas de gestionar la información de la Industria en cuanto a las Series históricas; por ello surge el presente trabajo de diploma en el marco de trabajo de las relaciones de colaboración entre la UCI y la ONEI; haciendo énfasis en la técnica de almacenamiento de datos utilizada.

1.3 Almacenes de datos

1.3.1 Definición de Almacenes de datos

Con la era de la informática surge simultáneamente el problema del almacenamiento de los datos para el hombre, debido al incremento de la cantidad de información que se generaba a diario se hacía más complejo su almacenamiento. Por tal razón, a partir de las atenuantes planteadas surgieron numerosas aplicaciones que fueron idóneas de aglutinar las informaciones capaces de aliviar la situación existente, como por ejemplo: Los AD, los cuales con el transcurso de los años y el avance de las tecnologías se han ido perfeccionando.

Existen diversos conceptos de los AD por lo que es muy difícil encontrar una concepción completa y

compartida por los autores, lo que demuestra que se trata de una herramienta en evolución y de compleja concepción.

"*Un AD es una gran colección de datos que recoge información de múltiples sistemas fuentes u operacionales dispersos, y cuya actividad se centra en la toma de decisiones. Una vez reunidos los datos de los sistemas fuentes se guardan durante mucho tiempo, lo que permite el acceso a datos históricos; así los AD proporcionan al usuario una interfaz consolidada única para los datos, lo que hace más fácil escribir las consultas para la toma de decisiones"*, así lo define Roberto Hernando Velasco. (1)

Según Kimball un AD *"es una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis"*, también determinó que un AD no era más que: *"la unión de todos los MD de una entidad".* (2)

Por otra parte Josep Curto en su blog lo aborda de la siguiente manera: *"proporciona una visión global, común e integrada de los datos de la organización, independiente de cómo se vayan a utilizar posteriormente por los consumidores o usuarios."* (3)

"AD es un conjunto de datos integrados, históricos, variantes en el tiempo y unidos alrededor de un tema específico, que es usado por la gerencia para la toma de decisiones", así lo defiende otro reconocido autor, Bill Inmon, el padre de los AD. (4)

Atendiendo a lo anteriormente expresado, son disímiles las definiciones proporcionadas por diversos autores en alguna ocasión, que de forma muy particular expresan lo que para ellos representa un AD. En la presente investigación se asume como AD una colección de datos orientados a temas específicos, en los que la información va a prevalecer de manera segura cierto tiempo sin importar de qué forma pueda ser utilizada posteriormente, siendo este un aspecto de gran relevancia en el proceso de toma de decisiones que se llevan a cabo dentro de los organismos y entidades presentes hoy día, como se ilustra en la **figura 1**.

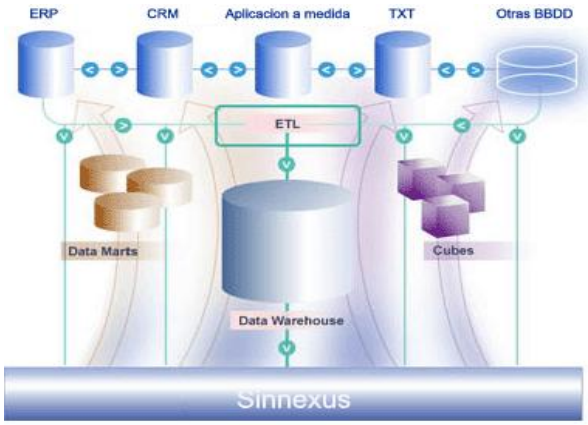


Figura 1. Almacén de datos.

1.3.2 Características de Almacenes de datos

- **Organizado en torno a temas.**

Significa que la información se clasifica en base a los aspectos que son de interés para la empresa. Siendo así, los datos tomados están en contraste con los clásicos procesos orientados a las aplicaciones, como se muestra en la **figura 2**.

Las diferencias entre la orientación de procesos y funciones de las aplicaciones y la orientación a temas, radican en el contenido a nivel de los datos. En el AD se excluye la información que no será utilizada por el sistema de soporte de decisiones, mientras que la información de las orientadas a las aplicaciones, contienen los datos para satisfacer de inmediato los requerimientos funcionales y de proceso, que pueden ser usados o no por el analista de soporte de decisiones.

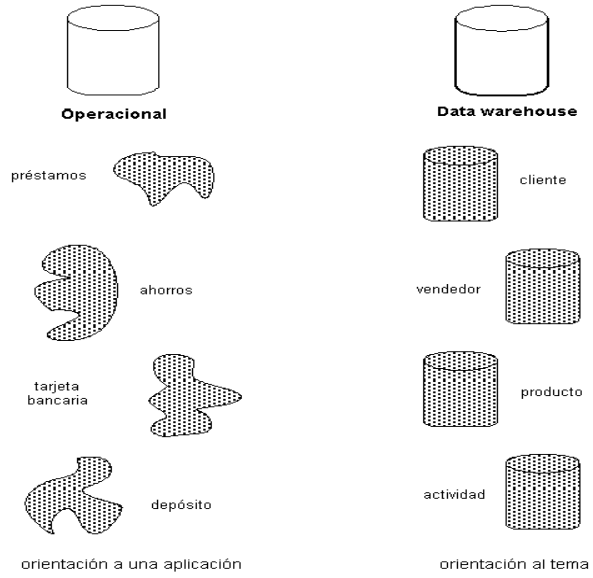


Figura 2. Orientado a temas.

- **Integrado**.

La integración de datos se muestra de muchas maneras: en convenciones de nombres consistentes, en la medida uniforme de variables, en la codificación de estructuras consistentes, en atributos físicos de los datos consistentes, fuentes múltiples y otros.

El contraste de la integración encontrada en el AD con la carencia de integración del ambiente de aplicaciones se muestra en la **figura 3**, con diferencias bien marcadas.

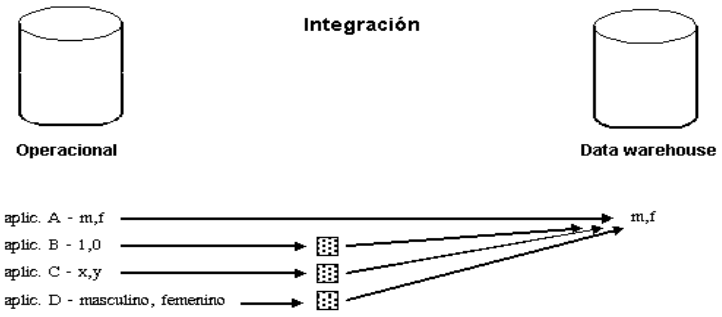


Figura 3. Integración de los datos.

- **Dependiente del tiempo**.

Toda la información del AD es requerida en algún momento. Esta característica básica de los datos en un depósito es muy diferente de la información encontrada en el ambiente operacional. En éstos, la información se requiere al momento de acceder. En otras palabras, en el ambiente operacional, cuando se acceda a una unidad de información, se espera que los valores requeridos se obtengan a partir del momento de acceso.

Como la información en el AD es solicitada en cualquier momento (es decir, no "ahora mismo"), los datos encontrados en el depósito se llaman de "tiempo variante".

Los datos históricos son de poco uso en el procesamiento operacional. La información del depósito por el contraste, debe incluir los datos históricos para usarse en la identificación y evaluación de tendencias, como se muestra en la **figura 4**.

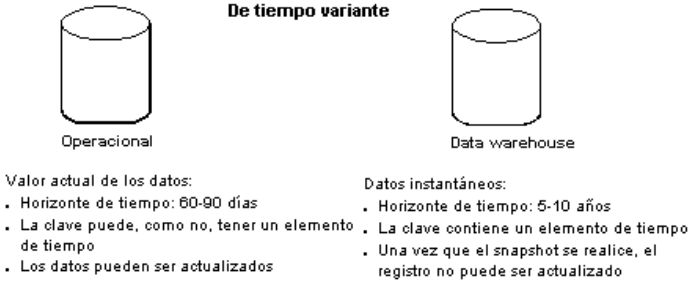


Figura 4. Variante en el tiempo.

- **No volátil**.

La información es útil sólo cuando es estable. Los datos operacionales cambian sobre una base momento a momento. La perspectiva más grande, esencial para el análisis y la toma de decisiones, requiere una base de datos estable.

En la **figura 5** se muestra que la actualización (insertar, borrar y modificar), se hace regularmente en el ambiente operacional sobre una base de registro por registro. Pero la manipulación básica de los datos que ocurre en el AD es mucho más simple. Hay dos únicos tipos de operaciones: la carga inicial de

datos y el acceso a los mismos. No hay actualización de datos (en el sentido general de actualización) en el depósito, como una parte normal de procesamiento.

Hay algunas consecuencias muy importantes de esta diferencia básica, entre el procesamiento operacional y del AD. En el nivel de diseño, la necesidad de ser precavido para actualizar las anomalías no es un factor en el AD, pues no se hace la actualización de datos. Esto significa que en el nivel físico de diseño, se pueden tomar libertades para optimizar el acceso a los datos, particularmente al usar la normalización y desnormalización física. (5)

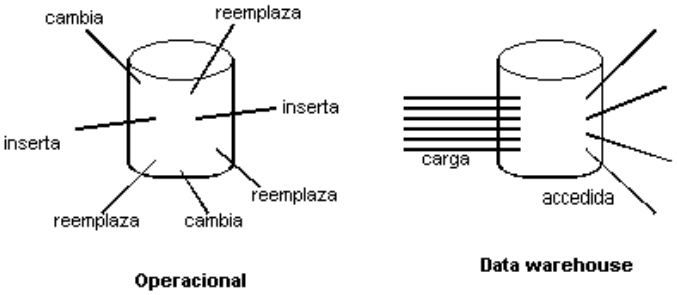


Figura 5. Información no volátil.

1.3.3 Componentes de los Almacenes de datos

✓ **Fuentes de datos:** este componente es el que normalmente está presente originariamente en las organizaciones, y a partir del cual se realiza la captura de datos que se contemplará en el AD. Estas fuentes de datos pueden ser sistemas operacionales corporativos (representan el entorno del que se obtienen la mayor parte de los datos significativos de la operativa diaria de la compañía), sistemas operacionales departamentales, fuentes externas, etc.

✓ **Extracción y transformación:** es responsable de que la información pueda moverse, con las trasformaciones que sean necesarias, desde las fuentes de datos antes mencionados al AD.

✓ **Servidor de datos:** también podría denominarse componente de gestión. Los servicios que debe ofrecer incluye un servicio de mantenimiento de datos y un servicio de distribución para exportar datos del AD a servidores de bases de datos descentralizados y otros sistemas de soporte de decisiones de usuario. El componente de gestión también ofrece servicios de seguridad (archivo, backup, recuperación) y monitorización. Generalmente estos servicios utilizan los medios suministrados por el software del sistema operativo y de bases de datos subyacente. El componente de SGBD (Sistema de Gestión de Bases de Datos) consiste en el software de base de datos que se utilice para mantener y extraer datos. Hay dos enfoques diferentes para el almacenamiento de la información: las bases de datos relacionales (SGBDR) o gestores de bases de datos multidimensionales (SGBDM).

✓ **Herramientas de acceso:** sin las herramientas adecuadas de acceso y análisis el AD se puede convertir en una amalgama de datos sin ninguna utilidad. Es necesario poseer técnicas que capturen los datos importantes de manera rápida y puedan ser analizados desde diferentes puntos de vista. También deben trasformar los datos capturados en información útil para el negocio. Actualmente a este tipo de herramienta se les conocen como herramientas de BI y están situadas conceptualmente sobre el AD. Cada usuario final debe seleccionar que herramienta se ajusta mejor a sus necesidades y a su AD. Entre ellas podemos citar las consultas SQL (Structured Query Language), las herramientas ROLAP (Relational On Line Analytical Processing) y las herramientas DATA MINIG, como se representa en la **figura 6**. (6)

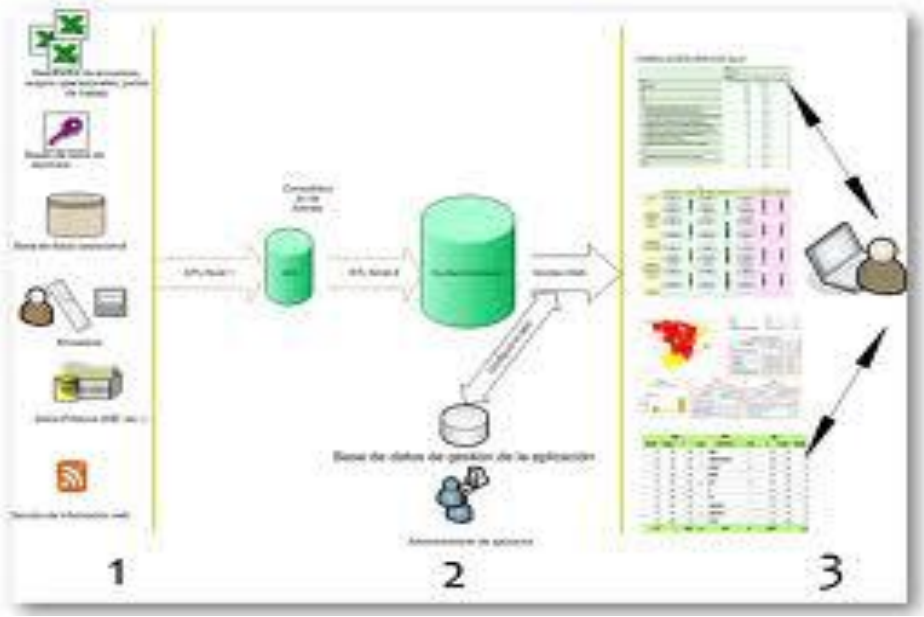


Figura 6. Componentes del Almacén de datos.

1.3.4 Ventajas y Desventajas de los Almacenes de datos

➢ **Ventajas**

- ✓ Obtiene respuestas en tiempos razonables.
- ✓ Les permite tener fuentes externas para ayudar a nuestra información.
- ✓ La información proveniente de fuentes operacionales es trasformada y limpiada para lograr consistencia.
- ✓ Mejoran la eficiencia en la toma de decisiones de una organización.
- ✓ Hacen más fácil el acceso a cierta cantidad de datos.
- ✓ Integra múltiples informaciones provenientes de varios sistemas externos. (7)

➢ **Desventajas**

1. Problemas con los sistemas de origen de los datos.

2. Pueden suponer altos gastos, además de los gastos de mantenimiento que son muy elevados.

3. Pueden quedarse obsoletos relativamente pronto si los usuarios incrementan sus necesidades.

4. La construcción de un AD puede requerir de mucho tiempo. (8)

1.3.5 Aportes de los Almacenes de datos

✓ Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio.

✓ Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor añadido para el negocio de dicha información.

✓ Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.

✓ Simplifica dentro de la empresa la implantación de sistemas de gestión integral de la relación con el cliente.

✓ Supone una optimización tecnológica y económica en entornos de Centro de Información, estadística o de generación de informes con retornos de la inversión espectaculares. (9)

1.3.6 Modelo Multidimensional

En los AD es más conveniente utilizar un modelo multidimensional (MMD) atendiendo que este posee sus ventajas con respecto al modelo entidad-relación (MER) aunque ambas almacenan la misma información, el MMD se representa a través de las tablas de hechos con sus concernientes tablas de dimensiones, como se muestra en la **figura 7**.

Figura 7. Modelo de datos multidimensional.

➢ **Tablas de Hechos**: Representan la ocurrencia de un determinado proceso dentro de la organización y no tienen relación entre sí. Generalmente, almacenan medidas numéricas, las que

representan valores de las dimensiones, aunque en ocasiones estas no están presentes y se les denominan “tablas de hechos sin hechos”, es decir, la relación entre las dimensiones que definen la llave de esta tabla de hecho implica por sí sola la ocurrencia de un evento. La llave de la tabla de hecho, es una llave compuesta, debido a que se forma de la composición de las llaves primarias de las tablas dimensionales a las que está unida, como se representa en la **figura 8**.

- **Tablas de Dimensiones**: Contienen, generalmente, una llave simple y atributos que la describen. En dependencia del esquema de diseño que se asuma pueden contener llaves foráneas de otras tablas de dimensión. Existe una dimensión fundamental en todo AD, la dimensión tiempo. Esto ocurre porque todo registro que se incluya constituye la ocurrencia de un fenómeno en un instante de tiempo definido. Dicha dimensión es la que establece uno de los objetivos fundamentales de la construcción de un AD, la conservación de un “histórico”. Los atributos dimensionales son fundamentalmente textos descriptivos, estos juegan un papel determinante porque son la fuente de gran parte de todas las necesidades que deben cubrirse, además, sirven de restricciones en la mayoría de las consultas que realizan los usuarios. Esto significa, que la calidad del modelo multidimensional, dependerá en gran parte de cuan descriptivos y manejables, sean los atributos dimensionales escogidos. La dimensión más importante de un AD, es *la dimensión tiempo*, atendiendo que esta será la encargada de decir en qué momento ocurrió este hecho, como se ilustra en la **figura 8**.

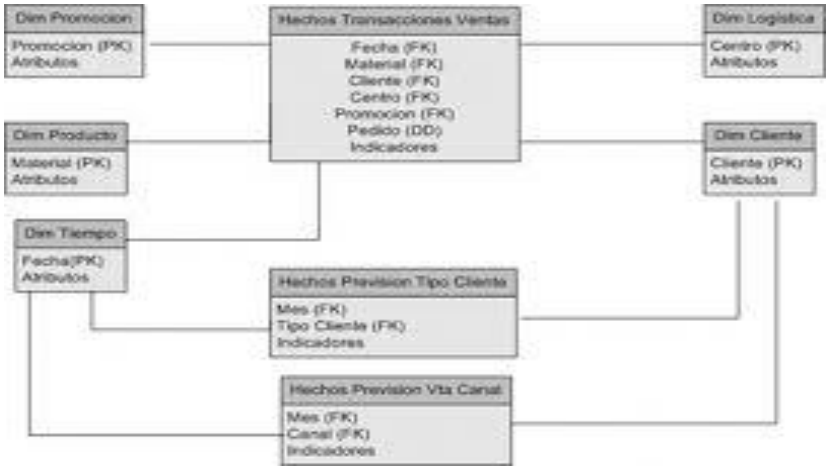


Figura 8. Tablas hechos y dimensiones de base de datos.

Existen varios esquemas para el modelado de los datos en un AD siendo los más utilizados:

- **Esquema de Estrella:** La tabla de hechos está en el centro de la estrella y están relacionadas con ella de forma radial todas las tablas de dimensiones, las cuales no se relacionan entre sí. No existen caminos alternativos en las dimensiones, como se representa en la **figura 9**.

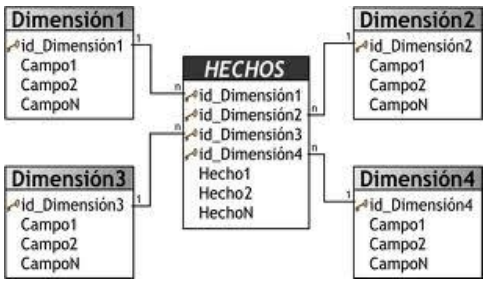


Figura 9. Esquema estrella.

➢ **Esquema de Copo de Nieve:** Es parecido al de estrella pero existen jerarquías en las dimensiones. Las tablas de dimensiones pueden estar relacionadas, o sea, existen caminos alternativos en ellas, como se muestra en la **figura 10**. (9)

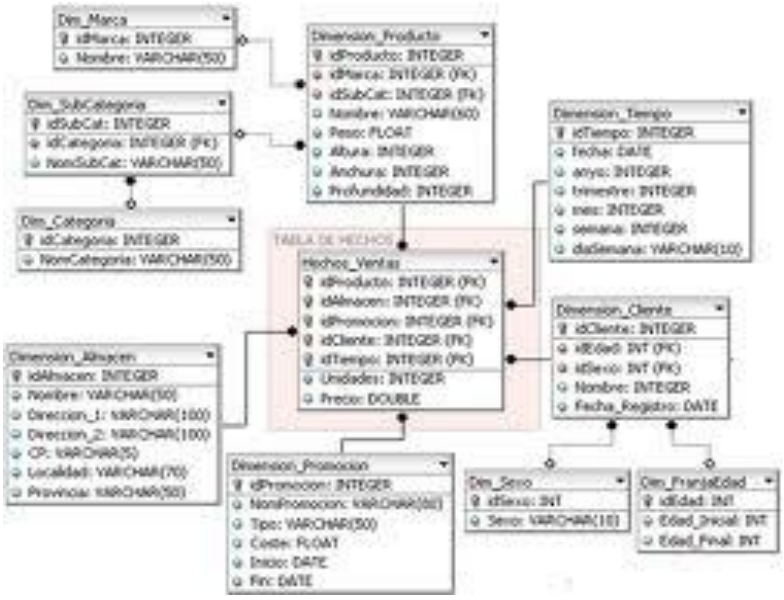


Figura 10. Esquema copo de nieve.

1.3.7 Modo de almacenamiento de datos OLAP

Las herramientas OLAP se basan en el modelo multidimensional de datos, presentando a los usuarios una visión multidimensional de los datos, independientemente del servidor que soporte el AD. Permiten un análisis multidimensional de las bases de datos de gran volumen para realizar un análisis especial de los mismos, más que a un análisis multidimensional, está destinada a mostrar cualquier correlación dentro de un volumen de datos importante del sistema de información con el fin de detectar alguna tendencia, como se ilustra en la **figura 11**.

Figura 11. Procesamiento analítico en línea

A continuación se mencionan algunas de sus características:

✓ La accesibilidad a los datos mayoritariamente es de sólo lectura. La gestión de los datos es poco común.

✓ Los datos se organizan según las áreas de negocio, y los formatos de los datos están integrados de manera uniforme en toda la organización.

✓ El historial de datos es a largo plazo, normalmente de hasta cinco años.

✓ Las bases de datos OLAP se suelen alimentar de información procedente de los sistemas operacionales existentes, mediante un proceso de extracción, transformación y carga (ETL).

Estas herramientas poseen un grupo de ventajas frente a las herramientas que se basan en el modelo del procesamiento transaccional en línea (OLTP), dentro del grupo de herramientas que se basan en este modelo podemos encontrar: archivos de texto, hojas de cálculo y las bases de datos transaccionales. Los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones. Una transacción genera un proceso atómico, y que puede involucrar operaciones de inserción, modificación y borrado de datos. El proceso transaccional es típico de las bases de datos operacionales.

El modo de almacenamiento de una partición afecta al rendimiento de las consultas y el procesamiento, a los requisitos de almacenamiento y a las ubicaciones de almacenamiento de la partición y de su grupo de medida y cubo primario. La elección del modo de almacenamiento afecta también a las opciones de procesamiento.

Una partición puede utilizar uno de estos tres modos de almacenamiento básicos:

- ♦ OLAP multidimensional (MOLAP)
- ♦ OLAP relacional (ROLAP)
- ♦ OLAP híbrido (HOLAP)

❖ **MOLAP (Procesamiento Analítico en Línea Multidimensional)**

El modo de almacenamiento MOLAP da lugar a que las agregaciones de la partición y una copia de sus datos de origen se almacenen en una estructura multidimensional. Esta estructura está muy

optimizada para maximizar el rendimiento de las consultas. La ubicación de almacenamiento puede estar en el equipo en donde se define la partición o en otro equipo. Dado que una copia de los datos de origen reside en la estructura multidimensional, las consultas se pueden resolver sin necesidad de obtener acceso a los datos de origen de la partición. Si se utilizan agregaciones, los tiempos de respuesta a las consultas pueden disminuir notablemente. Los datos de dicha estructura de la partición están tan actualizados como el procesamiento más reciente de la misma.

A medida que cambian los datos de origen, los objetos con almacenamiento MOLAP se deben procesar periódicamente para incorporar estos cambios y ponerlos a disposición de los usuarios. El procesamiento actualiza los datos en la estructura MOLAP, ya sea completamente o incrementalmente. Es posible actualizar los objetos en almacenamiento MOLAP completamente o incrementalmente sin desconectar la partición o el cubo. También se puede usar el almacenamiento en caché automático para minimizar la latencia y maximizar la disponibilidad, a la vez que se mantiene gran parte de las ventajas de rendimiento del almacenamiento MOLAP.

- **ROLAP (Procesamiento Analítico en Línea Relacional)**

El modo de almacenamiento ROLAP hace que las agregaciones de la partición se almacenen en vistas indizadas de la base de datos relacional que se especificó en el origen de datos de la partición. A diferencia del modo de almacenamiento MOLAP, ROLAP no hace que se almacene una copia de los datos del origen en las carpetas de datos. En su lugar, cuando no se pueden derivar los resultados de la caché de consultas, se utilizan las vistas indizadas del origen de datos para responder a las consultas. La respuesta a las consultas suele ser más lenta con el almacenamiento con tal estructura que con los modos de almacenamiento MOLAP o HOLAP. El tiempo de procesamiento también suele ser más lento con ROLAP. No obstante, la misma 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, como datos puramente históricos.

- **HOLAP (Procesamiento Analítico en Línea Híbrido)**

El modo de almacenamiento HOLAP combina atributos de los modos MOLAP y ROLAP. Al igual que MOLAP, HOLAP hace que las agregaciones de la partición se almacenen en una estructura multidimensional, no hace que se almacene una copia de los datos de origen. HOLAP es el equivalente de MOLAP para las consultas que sólo tienen acceso a los datos de resumen de las agregaciones de una partición. Con el modo de almacenamiento HOLAP, los usuarios suelen experimentar notables diferencias en cuanto a los tiempos de las consultas según si la consulta se puede resolver desde la caché o las agregaciones frente a los propios datos de origen.

Las particiones almacenadas como HOLAP son más pequeñas que sus equivalentes MOLAP dado que no contienen datos de origen y responden más rápidamente que las particiones ROLAP a las

consultas que implican datos de resumen. El modo de almacenamiento HOLAP suele ser más adecuado para particiones en cubos que requieren una respuesta de consultas rápida para resúmenes basados en una gran cantidad de datos de origen. No obstante, si los usuarios generan consultas que deben utilizar datos del nivel hoja (por ejemplo, para calcular valores medios), MOLAP suele ser una opción mejor. (10)

Después del estudio realizado en la presente investigación se decidió usar específicamente para el almacenamiento la arquitectura ROLAP, permitiendo almacenar la información en una Base de Datos (BD) relacional, además que el SGBD que se utilizará no modela datos multidimensionalmente. Tal arquitectura es capaz de usar datos pre-calculados si estos están disponibles, o de generar dinámicamente los resultados si es preciso. Este diseño accede directamente a la información del AD, y soporta técnicas de optimización de accesos para acelerar las consultas.

1.3.8 Mercado de datos

Un MD es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. Un MD puede ser alimentado desde los datos de AD o integrar por sí mismo un compendio de distintas fuentes de información, como se muestra en la **figura 12**.

Figura 12. Mercado de datos.

Por tanto, para crear el MD de un área funcional de la empresa es preciso encontrar la estructura óptima para el análisis de su información, estructura que puede estar montada sobre una base de datos OLTP, como el propio AD, o sobre una base de datos OLAP. La designación de una u otra dependerá de los datos, los requisitos y las características específicas de cada departamento. De esta forma se pueden plantear dos tipos de MD:

- **Mercado de datos OLAP (Online Analytical Processing)**

Se basan en los populares cubos OLAP que se construyen agregando, según los requisitos de cada área o departamento, las dimensiones y los indicadores necesarios de cada cubo relacional. El modo de creación, explotación y mantenimiento de los cubos OLAP es muy heterogéneo, en función de la herramienta final que se utilice.

❖ **Mercado de datos OLTP (Online Transactional Processing)**

Pueden basarse en un simple extracto del AD, no obstante, lo común es introducir mejoras en su rendimiento (las agregaciones y los filtrados suelen ser las operaciones más usuales) aprovechando las características particulares de cada área de la empresa. Las estructuras más comunes en este sentido son las tablas reporte, que vienen a ser tablas de facturas reducidas (que agregan las dimensiones oportunas), y las vistas materializadas, que se construyen con la misma estructura que las anteriores, pero con el objetivo de explotar la reescritura de consultas (aunque sólo es posibles en algunos SGBD avanzados, como Oracle).

Los MD presentan las siguientes características:

- ✓ Es un subconjunto de un AD existente.
- ✓ Optimizado para consultas específicas.
- ✓ Altamente sumarizado.
- ✓ Específicas funciones del negocio.
- ✓ Datos históricos.
- ✓ Orientada a un grupo de usuarios.

Los MD que están dotados con estas estructuras óptimas de análisis presentan las siguientes ventajas:

- ✓ Poco volumen de datos
- ✓ Mayor rapidez de consulta
- ✓ Consultas SQL y/o MDX sencillas
- ✓ Validación directa de la información
- ✓ Facilidad para la hostilización de los datos. (11)

1.4 Metodología de Almacenes de datos

El desarrollo de los AD ha contribuido a resolver problemas de manejo y uso adecuado de grandes fuentes de datos. Su utilización ha crecido en gran medida porque hacen más fácil el acceso de los usuarios finales a una gran variedad de información y pueden trabajar en conjunto, por lo tanto, aumentan el valor operacional de las aplicaciones empresariales, en especial la gestión de relaciones con clientes. Los AD se utilizan como una base de datos corporativa para posteriormente transformarlos en información útil para el usuario. Una de las vías para obtener un AD satisfactoriamente es con el uso de las metodologías.

Una metodología es una ciencia que estudia los métodos del conocimiento, amplios, complejos y transdisciplinaria con su objeto de estudio bien definido (los métodos), con normas o principios propios y una estructura. Tiene como objetivo el mejoramiento permanente de los procedimientos y criterios usados en la conducción de la indagación requerida para contestar preguntas y/o resolver problemas.

Estas se basan en la selección de técnicas a seguir para lograr un producto que permita alcanzar ciertos objetivos con calidad. (12)

Existen varias metodologías que se aplican a la creación de AD, algunas de ellas son las definidas por Inmon y Kimball; las mismas tienen sus propias características y pasos a seguir por lo que es de vital importancia escoger la que más se adecue al proyecto que se desea realizar.

En el desarrollo de la investigación se han analizado las metodologías mencionadas anteriormente, con el objetivo de profundizar en cada una y proponer la que esté acorde a las funciones del proyecto vigente, donde la misma permitirá establecer nuevas mejoras en su construcción, mayor rapidez, flexibilidad y sencillez en la implementación.

- **Bill Inmon**

Defiende una metodología descendente (top-down) a la hora de diseñar un AD, de esta forma se considerarán mejor todos los datos corporativos. En esta metodología los MD se crearán después de haber terminado el AD. Al tener este enfoque global, es más difícil de desarrollar en un proyecto sencillo (pues se intenta abordar el “todo”, a partir del cual luego se irá al “detalle”).

Inmon ve la necesidad de transferir la información de los diferentes OLTP de las organizaciones a un lugar centralizado donde los datos puedan ser utilizados para el análisis. La información ha de estar a los máximos niveles de detalle.

Esta metodología se caracteriza por ser:

- ✓ Orientado a temas: Los datos en la base de datos están organizados de manera que todos los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.
- ✓ Variante en el tiempo: Los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.
- ✓ No volátil: La información no se modifica, ni se elimina, una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.
- ✓ Integrado: La base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes. (4)

- **Ralph Kimball**

Se enfoca principalmente en el diseño de la base de datos que almacenará la información para el proceso de toma de decisiones. Su diseño se basa en la creación de tablas de hechos, es decir, tablas que contengan datos numéricos de los indicadores a analizar, o sea, los fundamentos cuantitativos de los datos. Las tablas anteriores se relacionan con tablas de dimensiones, las cuales contienen las características cualitativas de los indicadores, es decir, toda aquella información que clasifique los datos requeridos. A este modelo de datos se le conoce como diseño estrella, existen variaciones de éste llamados copo de nieve y diseño "flat". Todos estos diseños tienen la característica de preparar la

información de acuerdo a la necesidad de tomar decisiones y no a los argumentos técnicos de espacio de almacenamiento. Defiende por tanto una metodología ascendente (bottom-up) a la hora de diseñar un AD. (2)

1.4.1 Metodología a utilizar en el desarrollo del mercado de datos

Se decidió utilizar una metodología desarrollada por el DAD que toma como base la metodología de Kimball y se ajusta a las condiciones y características de la producción en DATEC y la UCI. La metodología organiza sus roles, actividades y artefactos teniendo en cuenta los cinco grupos de trabajos en los que está organizado el DAD y las fases del ciclo de vida.

Los grupos de trabajo son:

- ✓ **Grupo de Análisis:** En este grupo se realizan las actividades referentes a la Ingeniería de Requisitos. Sus miembros participan en el desarrollo del proyecto durante todo su ciclo de vida, aunque sus actividades fundamentales se centran en las fases de Levantamiento de requisitos, Diseño**,** y Prueba.
- ✓ **Grupo de Arquitectura y BD:** En este grupo se realizan todas las actividades relacionadas con la definición, diseño e implementación de arquitectura de la solución. También es este grupo quien se encarga de la implementación física y mantenimiento del repositorio de datos. Sus miembros participan desde las fases iníciales del proyecto para definir la arquitectura pero el volumen mayor de su trabajo corresponde a las fases de Arquitectura, Implementación y Despliegue.
- ✓ **Grupo de Extracción, Transformación y Carga (ETL):** Es el grupo encargado de realizar las tareas de ETL para integrar los datos existentes en las distintas fuentes y llevarlos hasta el AD. Las fases donde sus miembros participan de manera más activa son las de Diseño e Implementación y Despliegue la mayor carga de trabajo la tienen en los flujos de Arquitectura y Diseño, Implementación y Despliegue.
- ✓ **Grupo de Visualización de datos:** Este grupo realiza todas las actividades necesarias para presentar los datos que se guardan en el AD. Sus miembros participan en todas las fases del ciclo de vida del proyecto.
- ✓ **Grupo de Dirección:** En este grupo se realizan las tareas relacionadas con la gestión de proyecto. Además incluye personas que pueden ser externas al equipo del proyecto y que prestan servicios específicos en la fase de Prueba como son los especialistas de calidad. Los miembros de este laboran durante todo el ciclo de vida del proyecto y son los máximos responsables de la ejecución correcta del proyecto. (13)

Dentro del ciclo de vida de dicha metodología existen una serie de fases de trabajo mencionados a continuación:

✓ **Estudio preliminar y Planeación:** Se realiza un estudio minucioso en la entidad cliente. Esto incluye un diagnóstico de información, de datos y de infraestructura tecnológica, todo esto con el fin de determinar qué es lo que se desea construir y qué condiciones existen para el desarrollo y montaje de la misma. También se llevan a cabo las tareas de planeación del proyecto, se definen los objetivos, el alcance preliminar, los costos estimados, los recursos necesarios, y otras series de actividades.

✓ **Levantamiento de requisitos:** Se realiza en tres direcciones, 1ra. Identificación de las metas y objetivos de la organización, 2da. Identificación de las necesidades de información de los clientes y las reglas de negocio; y 3ra. Haciendo un levantamiento detallado de cada una de las fuentes de datos a integrar para validar la disponibilidad de la información. Es aquí donde se definen los requerimientos a partir de una comparación de las necesidades de información y las reglas del negocio con los elementos que realmente están disponibles en las fuentes. También se define la manera de presentar los datos según lo que desea el cliente y los requisitos técnicos de la solución.

Esta es una de las fases principales teniendo en cuenta que los AD se desarrollan para cubrir necesidades de información de las organizaciones y su mayor resultados es lograr satisfacer al cliente en este sentido. De su correcta realización dependerá el éxito del proyecto. Es importante señalar que los tiempos de duración de esta fase no deben extenderse desmesuradamente, convirtiéndose en algo tedioso y sin resultados. Para ello se necesita contar con un equipo bien preparado y con habilidades en el tema.

✓ **Arquitectura:** Se define la arquitectura de la solución según los requisitos no funcionales obtenidos. Es el momento donde se definen aspectos como, la seguridad del sistema, la comunicación entre los subsistemas, la tecnología a utilizar, hardware y software, entre otros aspectos de gran importancia. Vale aclarar que esta fase puede desarrollarse en paralelo con la fase de Levantamiento de requisitos, siempre y cuando los resultados del diagnóstico tecnológico realizado durante la fase de "Estudio preliminar" dejen bien definidas las características técnicas de la organización y el cliente sepa lo que desea.

✓ **Diseño e Implementación:** Se define el diseño de las estructuras de almacenamiento, se diseñan los procesos de integración de datos como: las reglas de extracción, transformación y carga, se diseñan los cubos para la presentación de los datos, así como el diseño visual de la aplicación definido por el cliente. Después se implementan cada uno de los subsistemas *(repositorio de datos, integración de datos, presentación de datos).* Se lleva a cabo el diseño físico del Repositorio de Datos, se crean las estructuras de almacenamiento con las particiones y agregaciones correspondientes. Se crea el Área Temporal de Almacenamiento, se ejecutan las reglas de extracción, transformación y carga, haciendo los ajustes para integrar la información. Se configuran e implementan las herramientas de Inteligencia de Negocio o *Business Intelligence* (BI) para obtener los reportes, gráficos, mapas y

otros que cubran los requerimientos firmados con el cliente.

La implementación puede realizarse de manera paralela en el caso del subsistema de integración de datos y de presentación de datos, lo que no ocurre con el repositorio de datos, debido a que ambos dependen de este. Esto no afecta para nada los tiempos de desarrollo, considerando que la implementación del repositorio de datos es relativamente corta y es la implementación de los procesos de integración de datos lo que más esfuerzo y tiempo requiere un una solución de este tipo.

✓ **Prueba:** Aquí se realizan varias pruebas, comenzando por las Pruebas de Unidad llevadas a cabo por los propios desarrolladores de cada uno de los grupos, luego las Pruebas de Integración y Sistema, hasta llegar a las Pruebas de Aceptación con el cliente final. Esta fase no es la única en la que se realizan pruebas durante el desarrollo del proyecto, en todas las fases hay actividades de aseguramiento de la calidad, pero sí constituye el momento ideal donde después de terminada la construcción de la solución se le pueden aplicar todo tipo de pruebas que certifiquen su calidad.

✓ **Despliegue:** Consta de dos etapas, la primera es un Despliegue Piloto, donde se configuran los servidores necesarios y se instalan las herramientas según la Arquitectura definida, se cargan una muestra de los datos en un ambiente controlado, con el fin de demostrarle al cliente final que la solución funciona. Durante esta etapa se realizan algunas de las pruebas propuestas en la fase anterior porque lo que tienden a coincidir en el tiempo. Una vez aceptada la solución por el cliente, se realiza la carga histórica de los datos, puede ser en el mismo entorno que el Despliegue Piloto u otro, todo depende de las condiciones del cliente. Es aquí el momento más idóneo para llevar a cabo la Capacitación y Transferencia Tecnológica a los clientes. El resultado fundamental es la solución desplegada en el entorno real y en correcto funcionamiento.

✓ **Soporte y Mantenimiento:** Comienza cuando la solución está implantada y en explotación, y se ejecuta según el contrato firmado y las condiciones de soporte establecidas. Puede realizarse a través de variados servicios, que pueden ser soporte en línea, vía telefónica, web, correo u otros, hasta el acompañamiento junto al cliente.

✓ **Gestión y administración del proyecto:** Esta fase se ejecuta a lo largo de todo el ciclo de vida del proyecto. Es aquí donde se controla, gestiona y chequea todo el desarrollo, los gastos, las utilidades, los recursos, las adquisiciones, los planes y cronogramas, entre otras actividades relacionadas con la gestión y administración de proyecto. Es esta fase la columna vertebrar del proyecto, si no se ejecutan de forma continua y correctamente las actividades que plantea es seguro el fracaso del proyecto. (13)

El desarrollo del MD realizado en la presente investigación, no transitará por todas las fases del ciclo de vida de la metodología a utilizar, solo llegará hasta la fase de pruebas. Atendiendo que las dos

fases posteriores a esta son ejecutadas por personal del Departamento, y la Gestión y Administración del proyecto es desarrollada por los especialistas que pertenecen al Grupo de dirección.

1.5 Herramientas de Almacenes de datos

1.5.1 Herramienta para el modelado de los datos

- **Visual Paradigm for UML 6.4**

Visual Paradigm para UML es una herramienta UML profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. El software de modelado UML ayuda a una más rápida construcción de aplicaciones de calidad, mejores y a un menor coste. Permite dibujar todos los tipos de diagramas de clases, código inverso, generar código desde diagramas y generar documentación.

Lista de características:

- ✓ Diagramas de Procesos de Negocio - Proceso, Decisión, Actor de negocio, Documento Modelado colaborativo con CVS y Subversion.
- ✓ Ingeniería de ida y vuelta.
- ✓ Ingeniería inversa - Código a modelo, código a diagrama.
- ✓ Ingeniería inversa Java, C++, Esquemas XML, XML, .NET exe/dll, CORBA IDL.
- ✓ Generación de código - Modelo a código, diagrama a código.
- ✓ Editor de Detalles de Casos de Uso.
- ✓ Generación de código y despliegue de EJB´s - Generación de beans para el desarrollo y despliegue de aplicaciones.
- ✓ Diagramas de flujo de datos
- ✓ Generación de bases de datos - Transformación de diagramas de Entidad-Relación en tablas de base de datos.
- ✓ Ingeniería inversa de bases de datos - Desde Sistemas Gestores de Bases de Datos (DBMS) existentes a diagramas de Entidad-Relación.
- ✓ Generador de informes para generación de documentación.
- ✓ Distribución automática de diagramas - Reorganización de las figuras y conectores de los diagramas UML.
- ✓ Importación y exportación de ficheros XMI.
- ✓ Editor de figuras. (14)

1.5.2 Sistema Gestor de Bases de Datos

- **PostgreSQL 9.1**

Es un sistema de gestión de bases de datos objeto-relacional que presenta prestaciones y funcionalidades similares a muchos gestores de bases de datos comerciales. Es uno de los más completo, atendiendo que permite métodos almacenados, restricciones de integridad, vistas, etc. Dicho sistema comprueba la integridad referencial con gran funcionalidad como base de datos, si bien un poco más lento que otros motores.

A continuación se enumeran algunas de sus características:

1. Cumple completamente con ACID *(Automaticidad, Consistencia, Aislamiento, Durabilidad)*
2. Cumple con ANSI SQL
3. Integridad referencial
4. Replicación (soluciones comerciales y no comerciales) que permiten la duplicación de bases de datos maestras en múltiples sitios de replica
5. Interfaces nativas para ODBC, JDBC, C, C++, PHP, Perl, TCL, ECPG, Python y Ruby
6. Reglas, Vistas, Triggers, Unicode, Secuencias, Herencia, Outer Joins, Sub-selects.
7. Procedimientos almacenados
8. Soporte nativo SSL
9. Lenguajes procedimentales
10. Bloqueo a nivel mejor que fila
11. Índices parciales y funcionales
12. Soporte para consultas con UNION, UNION ALL y EXCEPT
13. Herramientas para generar SQL portable para compartir con otros sistemas compatibles con SQL
14. Sistema de tipos de datos extensible para proveer tipos de datos definidos por el usuario, y rápido desarrollo de nuevos tipos
15. Funciones de compatibilidad para ayudar en la transición desde otros sistemas menos compatibles con SQL. (15)

- **Herramienta de administración**
  - **Pgadmin III 1.14.1**

Pgadmin es el más popular y característico de administración de código abierto, es una plataforma de desarrollo de PostgreSQL, la base de datos código abierto más avanzada del mundo. La aplicación se puede utilizar en Linux, FreeBSD, Solaris, Mac OSX y Windows.

Pgadmin está diseñado para responder a las necesidades de todos los usuarios, desde la escritura de

simples consultas SQL, hasta crear bases de datos complejas. La interfaz gráfica soporta todas las características de PostgreSQL y facilita la administración. La conexión con el servidor se puede hacer a través de TCP / IP o Unix Domain Sockets (en * nix), y puede ser encriptado SSL para la seguridad. No hay controladores adicionales, necesarios para comunicarse con el servidor de bases de datos.(16)
Seguidamente se mencionan algunas de sus principales características:

- ✓ Multiplataforma
- ✓ Diseñado para las últimas versiones de PostgreSQL.
- ✓ Ayuda en línea
- ✓ Interfaz multilingüe
- ✓ Acceso de datos
- ✓ Tener acceso a todos los objetos de PostgreSQL. (17)

1.5.3 Herramientas de Extracción, Transformación y Carga (ETL)

- ➢ **Pentaho Data Integrator 4.2.1**

Muchas organizaciones tienen información disponible en aplicaciones y base de datos separados. Pentaho Data Integration (PDI o kettle) extrae, limpia, transforma e integra tal información y la pone en manos del usuario, la cual contribuirá con el proceso de toma de decisiones. Provee una consistencia, una sola versión de todos los recursos de información, que es uno de los más grandes desafíos para las organizaciones de Tecnologías e Informática hoy en día. PDI permite una poderosa ETL (Extracción, Transformación y Carga). El uso de Kettle permite evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar.
La herramienta de integración de datos PDI fue concebida para apoyar el desarrollo de soluciones de BI mediante metodologías ágiles, reduciendo y optimizando el ciclo de vida de aplicaciones BI al permitir avanzar de forma paralela en el diseño de las ETL, modelamiento y visualización de datos, que a su vez ayuda a reducir costos, mejorar la productividad y acortar el tiempo necesario para obtener resultados concretos.
A parte de ser open source y sin costes de licencia, las características básicas de esta herramienta son:

- ✓ Entorno gráfico de desarrollo
- ✓ Uso de tecnologías estándar: Java, XML, JavaScript
- ✓ Fácil de instalar y configurar
- ✓ Multiplataforma: Windows, Macintosh, Linux
- ✓ Basado en dos tipos de objetos: Transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones)

- ✓ Incluye cuatro herramientas:
  - ➢ Spoon: para diseñar transformaciones ETL usando el entorno gráfico
  - ➢ PAN: para ejecutar transformaciones diseñadas con Spoon
  - ➢ CHEF: para crear trabajos
  - ➢ Kitchen: para ejecutar trabajos. (18)

Dicha herramienta permite desplegar soluciones que satisfacen inconvenientes de:

- ✓ Integración de datos mediante la creación de procesos ETL (*extracción, transformación y carga*).
- ✓ Procesos de consolidación de datos, flujos de trabajo y creación de AD.
- ✓ Visualización de datos mediante la creación de cubos OLAP y cuadros de mando.
- ✓ Generación de informes y análisis de KPI's.
- ✓ Procesamiento en línea de datos y la aplicación de técnicas de minería de datos.

- ➢ **Data Cleaner 1.5.4**

Es una aplicación de software libre para el análisis, perfiles, transformación y limpieza de datos. Estas actividades ayudan a administrar y controlar la calidad de los datos. Los datos de alta calidad son la clave para hacer que la información sea útil y aplicable a cualquier empresa moderna. Tal herramienta es la alternativa gratuita al software de gestión de datos maestros, metodologías, almacenamiento de datos, proyectos, la investigación estadística, la preparación para la extracción, transformación y carga (ETL) y más actividades.

DataCleaner está licenciado bajo los términos de la Licencia Pública General Menor (LGPL), que permite a cualquiera utilizar el software para todos los efectos, pero las modificaciones introducidas en el Código debe ser aportado a la comunidad. Normalmente, se utiliza una herramienta como esta antes, durante y después de cualquier actividad ETL.

- ✓ Antes, para profundizar en los orígenes de datos que está sobre el uso en su trabajo.
- ✓ Durante, si (cuando) se encuentra con cualquier desajustes inesperados durante el proceso de ETL.
- ✓ Después de asegurar la coherencia y la calidad en la fuente de datos que han poblado. (19)

1.5.4 Herramientas de Inteligencia de Negocio (BI)

- ➢ **Pentaho Schema Workbench 3.2.0**

Esta herramienta de la suite Pentaho tiene como objetivo facilitar la tarea de diseño de cubos OLAP. Su sencilla interfaz permite modelar un XML con el diseño del cubo a través de opciones lógicas e intuitivas que no requieren de un manejo avanzado de este formato de archivo.

Permite mejorar considerablemente los tiempos de desarrollo y desarrollar en la implementación de proyectos de soluciones analíticas.

Dentro de sus características destacamos:

- ✓ Diseñador intuitivo de esquemas OLAP.
- ✓ Permite crear, editar, actualizar y publicar esquemas OLAP para ser desplegados por aplicaciones de visualización de Pentaho.
- ✓ Acelera de manera considerable la construcción e implementación de este tipo de soluciones.(21)

- ➢ **Mondrian OLAP Server 3.8**

Online Analytical Processing es la tecnología que organiza la información en una estructura dimensional que proporcionará la posibilidad de moverse por la información desplazándose en sus dimensiones.

Mondrian es el motor OLAP de Pentaho que gestiona comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente. Aunque puede ser integrado independientemente en cualquier otra plataforma, y de hecho es el componente, junto con Data Integration que más se utiliza.

Esta herramienta es un motor HOLAP que combina la flexibilidad del los motores ROLAP con una caché que le proporciona velocidad.

A continuación se hace referencia a sus características:

- ✓ Es un motor ampliamente utilizado y consolidado en entornos JAVA
- ✓ Es el motor de la mayoría de soluciones de BI. Open Source. (20)

- ➢ **Apache Tomcat 6.6**

Es un servidor con un protocolo de transferencia de hipertexto (HTTP) y un objeto que recibe una solicitud y genera una respuesta sobre la base de esa petición (servlets). Es la implementación de referencias, las especificaciones de servlets (2.4) y de JSP (2.0). Es software libre (licencia Apache 2.0) gestionado por la fundación Apache. Puede funcionar como servidor HTTP o conectado a otro servidor HTTP como Apache HTTP Server o IIS. Puede ejecutar servicios web mediante Apache Axis. (22)

Tomcat funciona como un contenedor de servlets desarrollado bajo el proyecto Jakarta en la Apache Software Fundation. Tomcat implementa las especificaciones de los servlets y de JavaServer Pages (JSP) de Sun Microsystems. Dado que fue escrito en Java, funciona en cualquier sistema operativo que disponga de una máquina virtual Java. Es cada vez más utilizado por las empresas en los entornos de producción debido a su contrastada estabilidad. Resulta de gran utilidad para los programadores que deseen usar Tomcat como servidor Web autónomo, en entornos con alto nivel de

tráfico y alta disponibilidad. Constituye además una excelente herramienta para los principiantes. Esta herramienta puede ser utilizada por plataformas como: Windows, Linux, Mac OS X, Solaris, y FreeBSD, con sus ficheros de configuración específicos, y consejos paso a paso para implementar y correr aplicaciones Web eficazmente. (23)

- **Pentaho BI Server 3.6.0**

Es una aplicación 100% Java2EE que nos permite gestionar todos nuestros recursos de BI.
Cuenta con una Interfaz de Usuario de BI donde encontramos disponibles todos nuestros informes, vistas OLAP y cuadros de mando. Así mismo como el acceso a una consola de administración que permitirá gestionar y supervisar tanto la aplicación como los usuarios. Qué informes consulta cada usuario, cuándo se han consultado, el rendimiento de la aplicación, etc.
Presenta características tales como:

- ✓ Aplicación Java2EE 100% extensible, adaptable y configurable.
- ✓ La gestión de la configuración, tanto de la instalación inicial como del mantenimiento está muy bien resuelta.
- ✓ Se Integra con la mayoría de entornos y se puede comunicar con otras aplicaciones vía servicios web.
- ✓ Se integra con la mayoría de entornos y se puede comunicar con otras aplicaciones vía servicios web.
- ✓ Integra todos los recursos informacionales en una única plataforma de explotación
- ✓ Proporciona mucha libertad al usuario y los desarrolladores para crear contenidos nuevos.
- ✓ Explotación de sus recursos como SOAP servicios web. (20)

Conclusiones del capítulo

En el presente capítulo se analizaron los conceptos primordiales relacionados con los AD, además de sus componentes, las metodologías y herramientas que servirán para el desarrollo de una mejor solución del mercado de datos Series históricas de industria para el SIGOB, el cual contribuirá con el proceso de toma de decisiones para el área de Industria en la ONEI. Para el desarrollo de la presente investigación se escoge el modelo para el Desarrollo de Soluciones de AD e Inteligencia de Negocio propuesta por DATEC, la misma toma como base la metodología de Ralph Kimball, como SGBD PostgreSQL y como herramienta de modelado Visual Paradigm. Para guiar el proceso de ETL se eligieron las herramientas Pentaho Data Integration y Data Cleaner, mientras que para el desarrollo de BI el Pentaho Schema Workbench, Mondrian OLAP Server, Apache Tomcat y Pentaho BI Server.

Capítulo 2: Análisis y diseño del mercado de datos

Introducción

En el presente capítulo se realizará un estudio preliminar del negocio, se analizarán las necesidades de información, las reglas del negocio, la descripción de los casos de uso, la identificación de dimensiones, hechos y medidas, el desarrollo de la matriz BUS, el modelo de datos, la arquitectura de información, el diseño de los procesos de integración y de los reportes candidatos.

2.1 Análisis de la solución

El análisis es el proceso donde se entiende la estructura y el problema de la organización para la cual se desarrollará la aplicación. Además en él se identifican los requisitos del cliente, los requisitos funcionales, los no funcionales y las reglas del negocio, logrando una aproximación al diseño.

2.1.1 Descripción del negocio

La ONEI es el órgano rector en cuanto a estadística e información representa en Cuba. El mismo es el encargado de recoger y organizar toda la información proveniente de todo el territorio nacional, está compuesto por diferentes áreas dentro de las que se encuentra la de industria, la cual es la responsable de recopilar y analizar toda la información originada del sector económico industrial, dentro de la que aparecen las series históricas. Esta información se encuentra en formato excel y puede ser consultada en el sitio de la ONEI.

2.1.2 Tema de análisis identificado

Teniendo en cuenta el área de trabajo es que se define el tema de análisis. Para la construcción de la propuesta se definió como tema de análisis: Industria.

2.1.3 Reglas del negocio

Las reglas del negocio especifican las normas, políticas, restricciones y estructura de una organización. Las mismas pueden estar definidas en acuerdos o manuales de procedimiento. Es importante cumplir con ellas para alcanzar los objetivos de la organización. En la presente investigación se definieron cinco reglas del negocio, las cuales se encuentran descritas en el artefacto Reglas de negocio y transformación.

2.1.4 Necesidades de los usuarios

Es de vital importancia tener en cuenta las necesidades de los usuarios debido a que de esta forma se pueden crear productos con éxito. Estas constituyen la forma en que el usuario accede y analiza la información.

En las series, la información es almacenada anualmente y se recopila en cuanto al índice del volumen físico de la industria por el origen y el destino de los productos, los indicadores fundamentales de la industria azucarera, las producciones industriales seleccionadas y la dinámica de las producciones industriales seleccionadas.

2.1.5 Levantamiento de requisitos

Con la identificación de las necesidades de los usuarios, se definen los requisitos de información, los mismos describen que información debe almacenarse en el sistema para lograr satisfacer dichas necesidades. Estas son los reportes que el cliente necesita que se le muestren. Además se identifican los requisitos no funcionales que son capacidades o condiciones que el sistema debe cumplir y los requisitos funcionales que son propiedades o cualidades que el producto debe consumar.

En el mercado de datos Series históricas de industria se identificaron cinco requisitos de información, 25 requisitos funcionales y 21 requisitos no funcionales, los cuales se mencionan seguidamente:

Requisitos de información

- ✓ RI_1- Obtener el índice del volumen físico de la industria por origen de los productos y por año.
- ✓ RI_2- Obtener el índice del volumen físico de la industria por destino de los productos y por año.
- ✓ RI_3- Obtener los indicadores fundamentales de la industria azucarera por Zafra y por Concepto.
- ✓ RI_4- Obtener las producciones industriales seleccionadas por producciones y por año.
- ✓ RI_5- Obtener la dinámica de las producciones industriales seleccionadas por producciones y por año.

Requisitos funcionales

- ✓ RF_1- Autenticar usuario.
- ✓ RF_2- Adicionar usuarios.
- ✓ RF_3- Eliminar usuarios.
- ✓ RF_4- Adicionar roles.
- ✓ RF_5- Eliminar roles.
- ✓ RF_6- Insertar reportes.
- ✓ RF_7- Modificar reportes.
- ✓ RF_8- Eliminar reportes.

- ✓ RF_9- Extraer información.
- ✓ RF_10- Realizar transformación y carga.
- ✓ RF_11- Abrir navegador OLAP.
- ✓ RF_12- Mostrar editor MDX.
- ✓ RF_13- Mostrar Padres.
- ✓ RF_14- Ocultar repeticiones.
- ✓ RF_15- Intercambiar ejes.
- ✓ RF_16- Configurar impresión.
- ✓ RF_17- Exportar a PDF.
- ✓ RF_18- Exportar a excel.
- ✓ RF_19- Mostrar propiedades.
- ✓ RF_20- Suprimir filas.
- ✓ RF_21- Detallar miembros.
- ✓ RF_22- Entrar en detalles.
- ✓ RF_23- Mostrar datos de origen.

Requisitos no funcionales

- ✓ Cumplir con las pautas de diseño de las interfaces.

El sistema debe tener una interfaz gráfica uniforme que incluya pantallas, menús y opciones. Las pautas de diseño se realizarán siguiendo la arquitectura de información definida.

- ✓ Mostrar los mensajes, títulos y demás textos que aparezcan en la interfaz del sistema en idioma español.

Los títulos de los componentes de la interfaz, los mensajes para interactuar con los usuarios y los mensajes de error, deben ser en idioma español y tener una apariencia uniforme en todo el sistema. Los mensajes de error deberán ser lo suficientemente informativos para dar a conocer la severidad del error.

- ✓ Agilizar el acceso a los reportes del almacén de datos mediante la distribución de la información por áreas de análisis.

El usuario podrá acceder de manera rápida a la información que solicita en el área correspondiente de acuerdo al objetivo de su solicitud.

- ✓ Asegurar la disponibilidad del sistema.

El sistema debe estar disponible durante el horario de trabajo. En caso de fallo, la recuperación del servicio no deberá ser de un período de tiempo muy prolongado.

- ✓ Asegurar la recuperación ante un fallo.

El sistema debe ser capaz de recuperarse ante un fallo, teniendo en cuenta la complejidad y naturaleza de éste. El tiempo para su correcta recuperación fluctúa entre 10 minutos y 72 horas. Este tiempo comprende la solución al problema, así como su validación y prueba.

- ✓ Garantizar la persistencia de la información.

Se debe realizar un respaldo total de los datos del almacén de datos con una frecuencia anual. Esta información se almacenará en el edificio correspondiente a la oficina de estadísticas de La Habana y será responsabilidad del grupo de administración de redes de la ONEI.

- ✓ Lograr la homogeneidad de la estructura de los elementos definidos en el almacén.

Las estructuras del almacén de datos deben tener un nombre estándar teniendo en cuenta el tipo de estructura que sea.

- ✓ Utilizar los lenguajes de programación definidos durante la investigación.

Como lenguaje dentro del sistema gestor de base de datos para la programación en el almacén de datos se utilizará PL/pgSQL. En la implementación de los procesos de integración de datos se utilizará el lenguaje JavaScript. También se hará uso del lenguaje MDX para realizar las consultas.

- ✓ Utilizar el Sistema Gestor de Base de Datos definido durante la investigación.

El gestor de base de datos que se utilizará es PostgreSQL y como interfaz de administración de dicho gestor PgAdmin.

- ✓ Utilizar la herramienta de integración de datos definida durante la investigación.

Para el proceso de integración de datos se usará la herramienta Pentaho Data Integrator.

- ✓ Utilizar las herramientas para la implementación de la capa de inteligencia de negocios definidas durante la investigación.

De la suite Pentaho, se usarán los siguientes componentes:

- ♦ Schema Workbench: herramienta gráfica que se utiliza para construir el esquema multidimensional que soportará la creación de los reportes multidimensionales.
- ♦ Pentaho BI Server: servidor que se encarga de visualizar los reportes, tableros de control digital, controlar el acceso a la información y unificar en una solución de inteligencia de negocios el uso de las demás herramientas que componen la suite.
- ♦ Pentaho Administrator Console: herramienta para administrar el Pentaho BI Server, que permite la administración de las conexiones a las bases de datos, tareas programadas así como los roles y usuarios.

Para el uso de las herramientas anteriores se requiere la instalación de la máquina virtual de java (Java Virtual Machine 6.0).

- ✓ Confección de un manual de usuario.

El sistema debe estar acompañado de un documento que guiará la ejecución del usuario teniendo en cuenta cada funcionalidad.

- ✓ Acceso al sistema.

El usuario deberá acceder a la aplicación mediante el protocolo HTTP, usando preferiblemente el navegador web Firefox 2.0 en adelante.

- ✓ Garantizar una interfaz amigable al usuario.

El sistema debe tener una interfaz amigable y sencilla de utilizar, teniendo en cuenta que los usuarios finales no son personas adiestradas en el campo de la informática.

- ✓ Definir las interfaces de hardware que soportará el sistema.

El sistema podrá interactuar solamente con una interfaz de hardware: la impresora. Esta interacción se ocasionará cuando se necesite imprimir un reporte en formato físico. El acceso a la impresora será mediante el protocolo TCP/IP a través de la interfaz que ofrece el hardware.

- ✓ Proporcionar características mínimas de hardware a las estaciones de trabajo.

Características de un cliente ligero.

- ✓ Proporcionar características mínimas de hardware a los servidores.

Para lograr una explotación aceptable del sistema los servidores deben contar con los siguientes requerimientos de hardware:

- ♦ Windows server 2003.
- ♦ 1 GB RAM.
- ♦ 1 Microprocesador Core2Duo.

- ✓ Instalar en las estaciones de trabajo el software necesario para el correcto funcionamiento del sistema.

Las configuraciones de software de las máquinas clientes deben contar al menos con:

- ♦ Firefox 2.0 o superior.
- ♦ Java Virtual Machine 6.0 y Schema Workbench 3.2.1 en caso de que un usuario capacitado requiera la construcción de esquemas multidimensionales para el diseño de nuevos reportes.

- ✓ Entregar el sistema a la ONEI.

El sistema debe ser transferido a la ONEI mediante un proceso de transferencia una vez que esté en explotación, incluyendo el código fuente y la documentación correspondiente.

- ✓ Requerimientos legales, de derecho de autor y otros.

No se hace solicitud de derecho de autor, patentes, marca comercial o complacencia con logotipo para el software, debido a que se usan soluciones con Licencia Pública General (GNU GPL por sus siglas en inglés), bajo el principio de software libre.

2.1.6 Reportes candidatos

A partir de los requisitos de información anteriormente descritos, se definieron cinco reportes candidatos, los cuales se referencian a continuación:

- ✓ Índice del volumen físico de la industria por el origen de los productos.
- ✓ Índice del volumen físico de la industria por destino de los productos.
- ✓ Indicadores fundamentales de la industria azucarera.
- ✓ Producciones industriales seleccionadas.
- ✓ Dinámica de las producciones industriales seleccionadas.

Para mayor información consultar el artefacto reportes candidatos donde se encuentran descritos.

2.1.7 Casos de uso del sistema (CUS)

Los CUS forman parte del análisis y ayudan a describir lo que debe hacer el sistema para cumplir con las necesidades del cliente. Es decir, mediante estos se describe la interacción entre el usuario y el sistema. En la creación del diagrama de casos de uso del sistema (DCUS), quedaron agrupados en 11 casos de uso los 25 requisitos funcionales y los cinco requisitos de información, el cual se muestra a continuación conjuntamente con la descripción de cada uno de sus componentes.

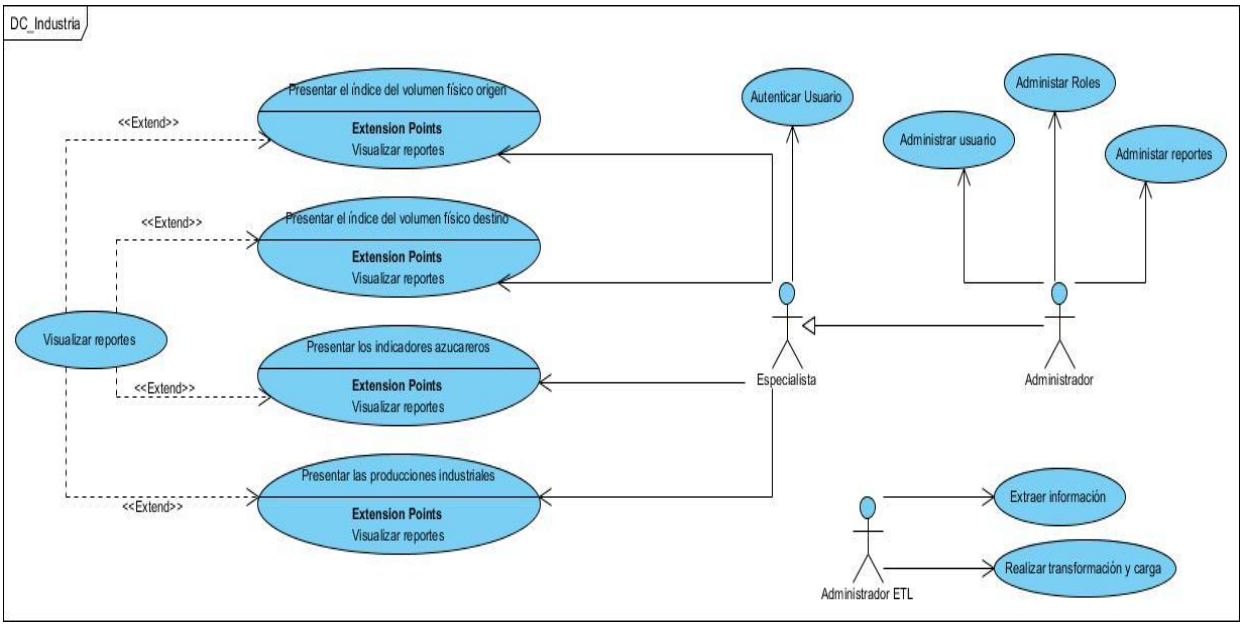


Figura 13. Diagrama de casos de uso.

Actores	Descripción
Especialista	Es el que tiene la tarea de realizar todos los pedidos de información que el cliente quiere que se muestren

Administrador	Es el encargado de administrar todo lo referente a usuarios, roles y reportes en el AD.
Administrador ETL	Es el encargado de realizar la extracción de la información desde las distintas fuentes, la transformación y carga de los datos al AD.

Tabla 1. Descripción de los actores

Casos de uso	Descripción
Presentar el índice del volumen físico origen.	Muestra los reportes referentes al índice del volumen físico de la industria por el origen de los productos.
Presentar el índice del volumen físico destino.	Muestra los reportes referentes al índice del volumen físico de la industria por el destino de los productos.
Presentar los indicadores azucareros.	Muestra los reportes referentes a los indicadores fundamentales de la industria azucarera.
Presentar las producciones industriales.	Muestra los reportes referentes a las producciones industriales seleccionadas y a la dinámica de dichas producciones.
Autenticar usuario	Este caso de uso permite que el usuario entre su usuario y contraseña para identificarse en el sistema y acceder al mismo, con los permisos asignados en dependencia del usuario.
Administrar usuario	Este caso de uso permite eliminar o adicionar un usuario al sistema.
Administrar reporte	Este caso de uso permite eliminar, modificar o adicionar un reporte al sistema.
Administrar roles	Este caso de uso permite eliminar o adicionar un rol.
Extraer datos	Este caso de uso permite realizar la extracción de los datos.
Realizar transformación y carga	Este caso de uso permite realizar las transformaciones y cargas a los datos.

Visualizar reportes	Este caso de uso permite seleccionar la opción que desea ejecutar y así hacerle cambios a los reportes.

Tabla 2. Descripción de Casos de uso

Consecutivamente se describe el casos de uso (CU) Presentar los indicadores azucareros, en el artefacto Especificación de casos de uso se describen el resto de los CU.

Objetivo	Presentar los indicadores azucareros.	
Actores	Especialista, Administrador	
Resumen	El caso de uso (CU) inicia cuando el especialista quiere analizar información referente a los indicadores fundamentales de la industria azucarera. El actor selecciona los reportes que desea analizar y se muestran los datos de los indicadores fundamentales de la industria azucarera. El caso de uso finaliza cuando el especialista termina de analizar la información correspondiente.	
Complejidad	Alta	
Prioridad	Media	
Precondiciones	El especialista tiene que estar autenticado. El MD tiene que estar poblado. Los reportes relacionados con la información correspondiente deben estar creados.	
Postcondiciones	Los reportes correspondientes fueron estudiados por el especialista.	
Flujo de eventos		
Flujo básico Presentar los indicadores azucareros.		
	Actor	**Sistema**
1	Selecciona el AAG-SIGOB.	
2		Muestra las áreas de análisis (AA) que están comprendidas dentro del AAG.
3	Selecciona el AA-Series industria	
4		Muestra los libros de trabajo (LT) que se corresponden con esa AA.
5	Selecciona el LT que desee.	
6		Muestra los reportes que contiene el LT seleccionado

7	Selecciona el reporte que desee.	
8		Visualiza el reporte seleccionado. Se ofrece las opciones que aparecen en la barra de herramienta. Ir al CU Visualizar reportes.
Flujos alternos		
Nº 2 Los datos son incorrectos.		
	Actor	**Sistema**
1		Muestra mensaje ¨ Datos de acceso incorrecto. Intente de nuevo¨

Opciones del reporte Presentar los indicadores azucareros.		
Perspectivas de análisis	**Posibles resultados**	
	Medidas	**Periodicidad**
3 Año-Zafra 4 Indicadores azucareros	Indicadores de la industria azucarera.	➢ Anual
Prototipo de interfaz de usuario		

Relaciones	**CU Incluidos**	No aplica
	CU Extendidos	CU Visualizar Reporte.
Requisitos no funcionales	Chequear el epígrafe **3.2 Requisitos no Funcionales** en el artefacto "**Especificación de requisitos**", los siguientes RnF 1, 2, 3, 4, 5, 6, 7, 14, 15, 16, 17, 20.	
Asuntos	Efectuar un perfeccionamiento en el CU.	

Tabla 3. Descripción del CU-Presentar los indicadores azucareros

2.1.8 Arquitectura de la solución

El mercado de datos Series históricas de industria cuenta con tres subsistemas, el de integración de datos, el de almacenamiento y el de visualización de la información.

✓ En el subsistema de integración de datos es donde se realizan todos los procesos ETL en los cuales se extrae, se limpia e integra toda la información almacenada en los sistemas fuentes a través de transformaciones y trabajos.

✓ En el subsistema de almacenamiento es donde se guarda toda la información que ha sido trasformada en el subsistema anterior.

✓ En el subsistema de visualización de la información es donde se muestra toda la información almacenada al cliente, a través de reportes. Los mismos permiten al cliente realizar un análisis de toda

la información procesada.

A continuación se representa la relación existente entre los componentes a desarrollar:

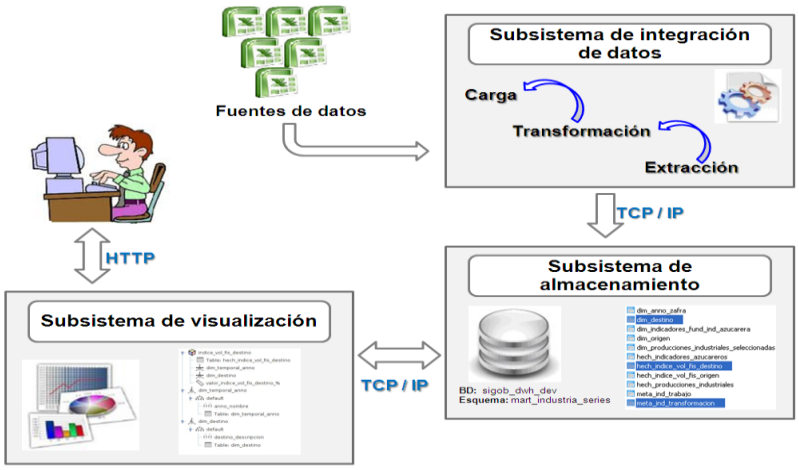


Figura 14. Arquitectura del sistema

2.1.9 Perfilado de datos

Con el perfilado de datos se logra una mayor comprensión de los mismos y mediante estos se definen nuevas reglas del negocio. Estas serán algunas de las reglas a utilizar en el proceso de ETL, el resto se encuentran en el artefacto “Reglas del negocio y transformación”, seguidamente se especifican:

- RN 1. El código de los atributos en cada una de las dimensiones no puede tomar valores repetidos.
- RN 2. En la fuente 11.1 correspondiente al Índice del volumen físico de la industria por el origen de los productos aparecen nulos los siguientes indicadores en los años especificados:
  - ✓ Fabricación de productos farmacéuticos y productos botánicos (de 1990 a 2001, 2009, 2010).
  - ✓ Fabricación de maquinarias y equipos (2007).
  - ✓ Suministro de gas (de 1990 a 2003, 2010).
  - ✓ Suministro de agua (de 2002 a 2010).
  - ✓ Suministro de electricidad (de 2002 a 2010).

2.2 Diseño

2.2.1 Matriz BUS o matriz dimensional

Es la representación de las relaciones que existen entre las tablas hechos y sus respectivas tablas de dimensiones. Posterior al análisis fueron identificados seis dimensiones y cuatro hechos, atendiendo a

que la información que se encuentra contenida en cada uno de ellos es diferente. Seguidamente se describen:

✓ **D1-** Dimensión temporal año: En la misma se recopilan todos los años comprendidos desde 1800 hasta 2020.

✓ **D2-** Dimensión origen: Dicha tabla almacenará los indicadores correspondientes al origen de los productos.

✓ **D3-** Dimensión destino: En esta tabla se guardan los indicadores referentes al destino de los productos.

✓ **D4-** Dimensión año zafra: La misma recoge la información correspondiente a los años de las zafras azucareras.

✓ **D5-** Dimensión indicadores fundamentales de la industria azucarera: En esta tabla se almacenarán los datos relacionados con los indicadores fundamentales de la industria azucarera. Dicha dimensión presenta una jerarquía de dos niveles.

✓ **D6-** Dimensión producciones industriales seleccionadas: En la presente tabla se recopilará la información correspondiente a las producciones industriales seleccionadas, presentando una jerarquía de cuatro niveles.

✓ **H1-** Hecho índice del volumen físico origen: En el mismo se recoge la información referente al índice del volumen físico por origen de los productos, la misma se da en por ciento (%).

✓ **H2-** Hecho índice del volumen físico destino: En esta tabla se recopilan los datos relacionados con el índice del volumen físico por el destino de los productos, la misma se da en por ciento (%).

✓ **H3-** Hecho indicadores azucareros: En el presente hecho se recogen los datos relacionados con los indicadores fundamentales de la industria azucarera, los cuales tienen sus propias unidades de medida.

✓ **H4-** Hecho producciones industriales: En el mismo se guardan datos relacionados con las producciones industriales, las mismas tienen sus propias unidades de medida y la dinámica de dichas producciones, las cuales se dan en por ciento (%).

Hechos	Dimensiones					
	D1	D2	D3	D4	D5	D6
H1	x	x				
H2	x		x			
H3				x	x	
H4	x					x

Tabla 4. Matriz BUS

2.2.2 Modelo de datos

A continuación se ilustra el modelo de datos dimensional. El mismo está compuesto por las tablas de hechos con sus respectivas medidas y tablas de dimensiones.

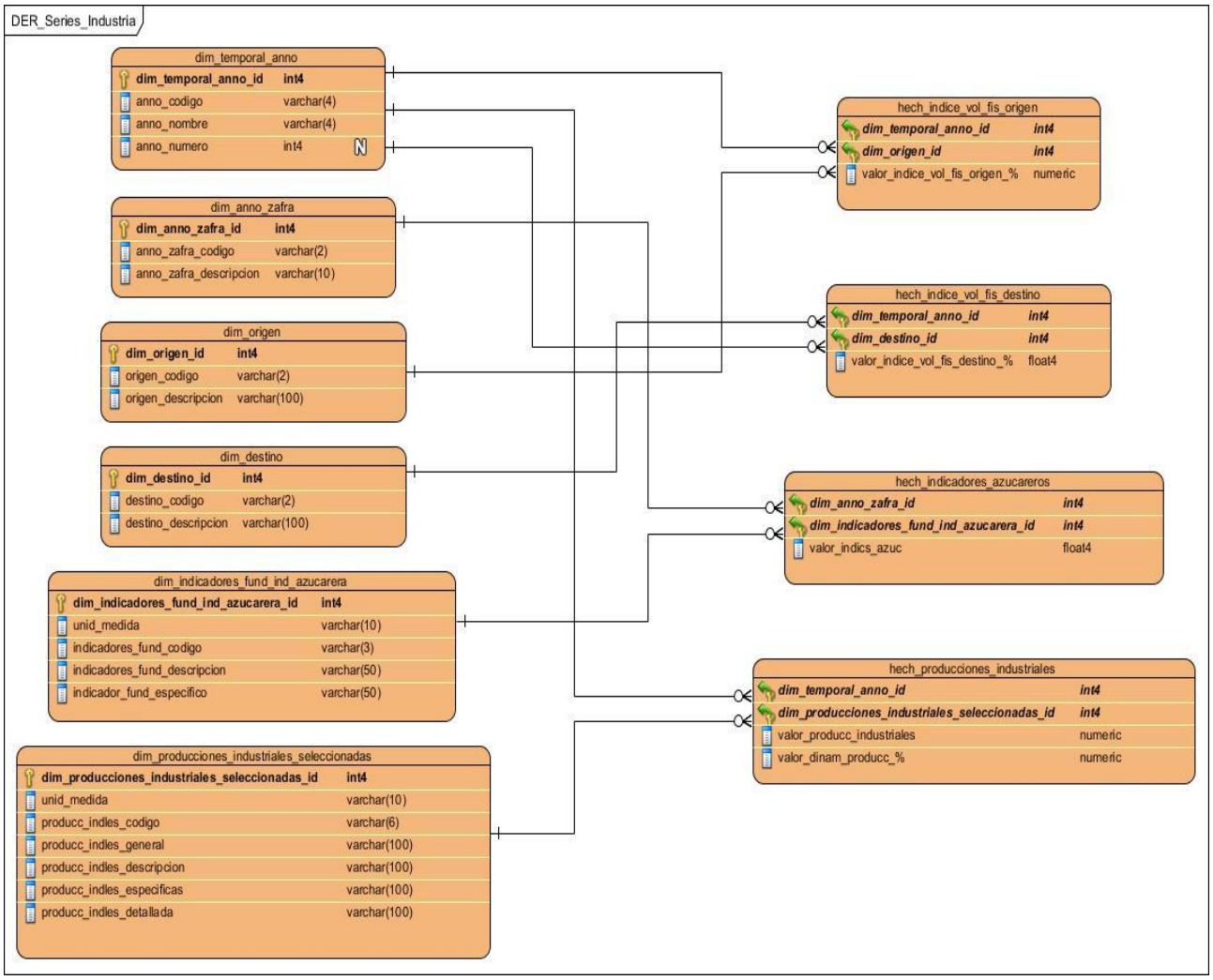


Figura 15. Modelo de datos dimensional.

Dimensiones

- ✓ **dim_temporal_anno:** La misma hace referencia a la variable de entrada tiempo.
- ✓ **dim_origen:** La misma hace referencia a la variable de entrada origen.
- ✓ **dim_destino:** La misma hace referencia a la variable de entrada destino.
- ✓ **dim_anno_zafra:** La misma hace referencia a la variable de entrada año-zafra.
- ✓ **dim_indicadores_fund_ind_azucarera:** La misma hace referencia a la variable de entrada indicadores azucareros.

✓ **dim_producciones_industriales_seleccionadas:** La misma hace referencia a la variable de entrada producciones industriales.

Hechos y medidas

✓ **hech_indice_vol_fis_origen:** El mismo hace referencia al hecho índice del volumen físico origen.

Medida:

→ valor_indice_vol_fis_origen (%): La misma hace referencia a la variable de salida valor del índice del volumen físico origen.

✓ **hech_indice_vol_fis_destino:** El mismo hace referencia al hecho índice del volumen físico destino.

Medida:

→ valor_indice_vol_fis_destino (%): La misma hace referencia a la variable de salida valor del índice del volumen físico destino.

✓ **hech_indicadores_azucareros:** El mismo hace referencia al hecho indicadores azucareros.

Medida:

→ valor_indics_azuc: La misma hace referencia a la variable de salida valor del indicador azucarero.

✓ **hech_producciones_industriales:** El mismo hace referencia al hecho producciones industriales.

Medida

→ valor_producc_industriales: La misma hace referencia a la variable de salida valor de las producciones industriales seleccionadas.

→ valor_dinam_producc (%): La misma hace referencia a la variable de salida valor de la dinámica de las producciones industriales.

2.2.3 Mapa de navegación

La información que se almacenará referente al mercado de datos Series históricas de industria estará conformada por un área de análisis. En la misma se agrupara la información brindada por el área de Industria. Contará con 4 libros de trabajos y estos son los encargados de agrupar todos los reportes generados en el área de análisis. Además tendrá los reportes que se componen cada uno de los libros de trabajo, siendo estos los que dan respuestas a las necesidades de información de los clientes. Seguidamente se ilustra como quedaría integrada al área de análisis general SIGOB el área de análisis con cada uno de sus libros de trabajo y estos con sus reportes.

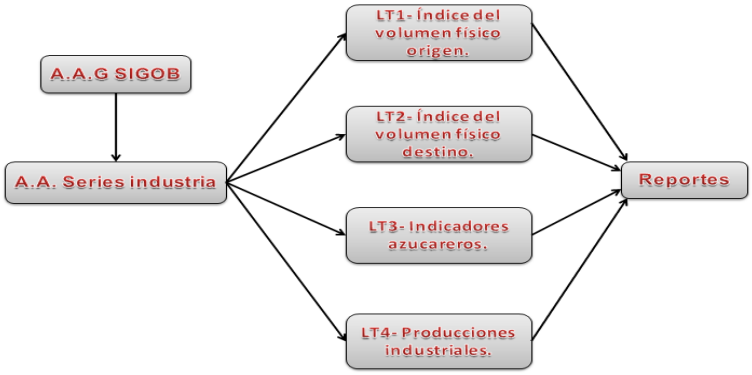


Figura 16. Diseño del mapa de navegación.

2.2.4 Diseño de los cubos OLAP

En el desarrollo del MD se diseñaron cuatro cubos multidimensionales: índice del volumen físico origen, índice del volumen físico destino, indicadores azucareros y producciones industriales. Para el diseño de los cubos OLAP se utilizó la herramienta Pentaho Schema Workbench, la cual facilita la creación de los cubos. Seguidamente se muestra el cubo multidimensional correspondiente al índice del volumen físico destino en la **figuras 17**:

Figura 17. Cubo índice volumen físico destino

Como se observa en las ilustraciones anteriores cada uno de los cubos se corresponde con hecho diferente, lo que significa que se definió un cubo para cada una de las tablas de hechos existentes en el mercado.

2.2.5 Diseño del subsistema de integración de datos

Para ordenar el proceso de ETL se diseñó como sería la realización del mismo. En la siguiente figura se muestra el flujo general del proceso.

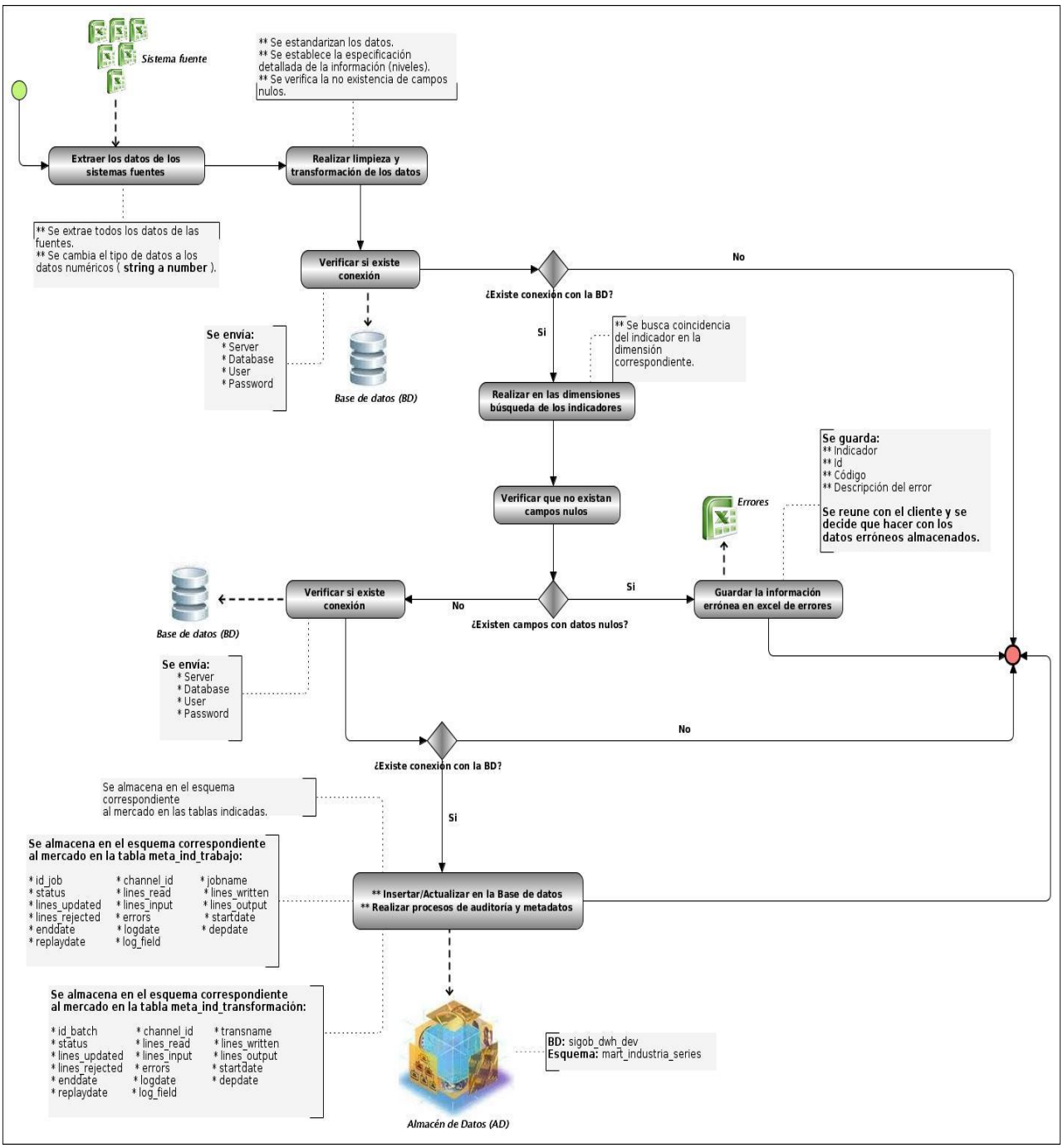


Figura 18. Diseño del proceso de ETL

Con el diseño de las transformaciones se describen los pasos que se deben seguir para realizar la carga de los hechos y las dimensiones al MD. Seguidamente se ilustra el diseño del flujo para llenar la dimensión destino.

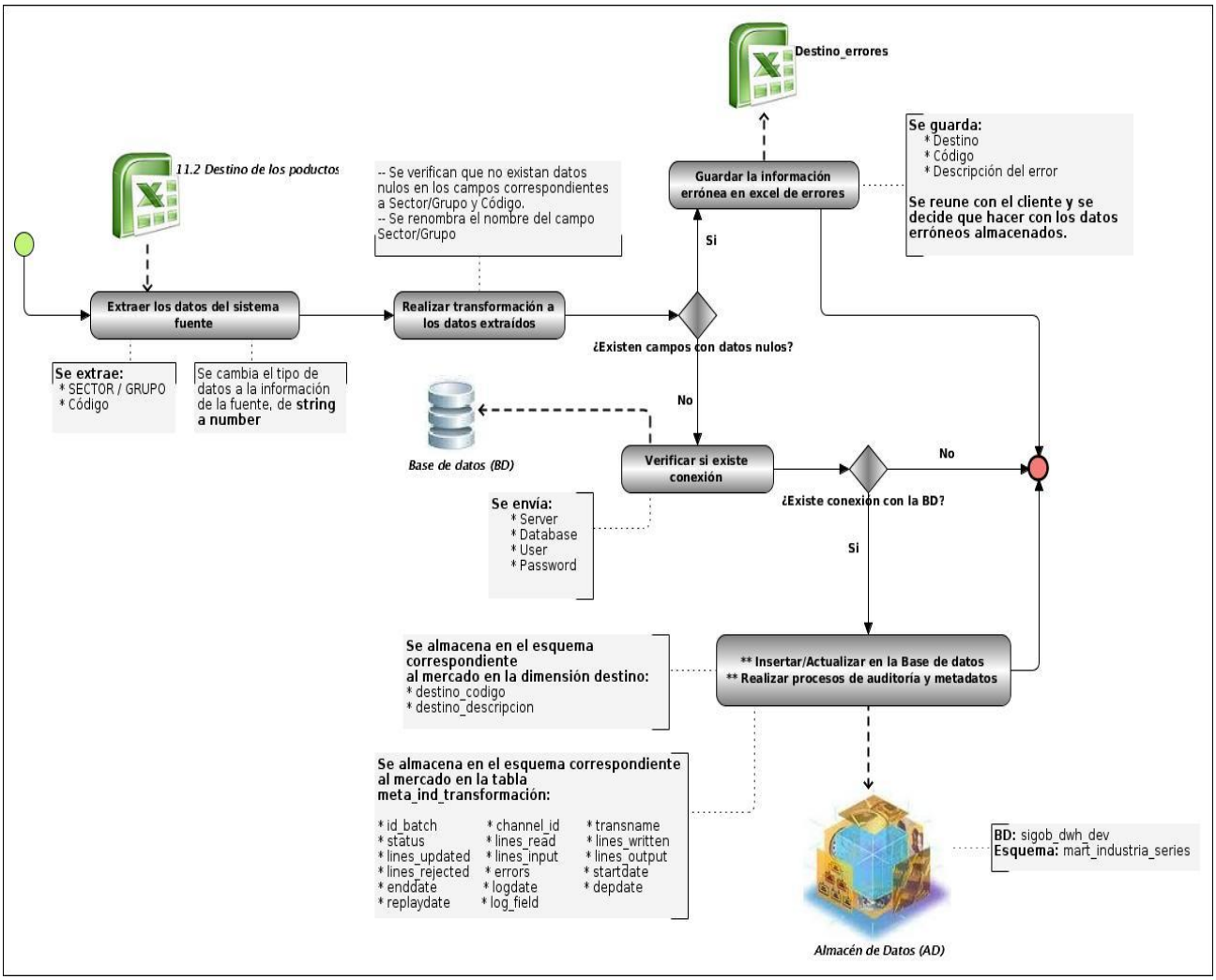


Figura 19. Diseño para llenar la dimensión destino

2.2.6 Esquema de seguridad

Los esquemas de seguridad no son más que los niveles de acceso, teniendo en cuenta los roles definidos. Seguidamente se muestran los roles y permisos que los usuarios poseen en su interacción con la base de datos y la aplicación.

Seguridad en la base de datos

Seguidamente se muestran los roles definidos para intercambio de los usuarios con la BD:

Usuarios	Descripción
Administrador	Está autorizado para eliminar, modificar e insertar en la base de datos.
Administrador de ETL	Efectúa los procesos de ETL de los datos.

Tabla 5. Usuarios de la base de datos

Seguridad en la aplicación

El acceso a los servicios del sistema deberá ser a través de un estándar web (http). El sistema permitirá la transmisión de la información por canales cifrados usando el protocolo HTTP. Las contraseñas de los usuarios no deben ser almacenadas en texto plano, esta información debe ser privada y específica de cada uno, de manera que nadie pueda reemplazar la identidad de un usuario en el sistema.

Las aplicaciones desplegadas en el servidor de BI de Pentaho muestran un continuo incremento, así como los usuarios que tienen acceso a estas. A continuación se definen los roles de seguridad para la interacción de los usuarios con la aplicación para hacer sostenible el manejo de la misma en este servidor.

Elementos de la aplicación	Roles con acceso
Área de análisis general (AAG)	Administrador Especialista
Carpeta raíz: AA. Series industria	Administrador Especialista

Tabla 6. Elementos y roles de acceso a la aplicación

Rol	Permisos
Administrador	Tiene acceso completo a todas las AAG.
Especialista	Tiene acceso de solo lectura al AA. Series industria. Visualiza los reportes.

Tabla 7. Roles y permisos

2.2.7 Políticas de respaldo y recuperación de datos

Estas son las medidas que se deben tener en cuenta en caso de ocurrencia de algún fallo en el sistema. Las que se pondrán en práctica en el MD son las siguientes:

Periodicidad de las salvas del sistema: Estas se harán anualmente a toda la información contenida en el mercado de datos Series históricas de industria, verificando que exista una copia de toda la información que haya sido guardada.

Tablas involucradas: Las tablas involucradas en la realización de las salvas son las que a continuación se mencionan:

- ✓ dim_temporal_anno
- ✓ dim_anno_zafra

- ✓ dim_indice_vol_fis_destino
- ✓ dim_indicadores_fund_ind_azucarera
- ✓ dim_producciones_industriales_seleccionadas
- ✓ hech_indice_vol_fis_origen
- ✓ hech_indice_vol_fis_destino
- ✓ hech_indicadores_azucareros
- ✓ hech_producciones_industriales

Conclusiones del capítulo

Se concluyó el capítulo con un estudio del negocio, alcanzando con esto mayores conocimientos sobre el flujo de información en el área de las Series históricas de industria. Se realizó un análisis profundo y detallado de los datos para la extracción de los mismos, lo cual ayudará al correcto avance en el MD. También quedaron identificados los requisitos funcionales, no funcionales e informativos, además de las reglas del negocio, para de esta forma cumplir con las necesidades del cliente. Para lograr un mejor entendimiento del negocio se creó el diagrama de casos de uso y el modelo de datos dimensional. Luego de terminado este capítulo quedan creadas las bases para entrar en la implementación del MD.

Capítulo 3. Implementación del mercado de datos

Introducción

En el mismo se implementará el proceso de extracción, transformación y carga (ETL), y la capa de inteligencia del negocio para el área de la Industria en la ONEI, específicamente las Series históricas de dicha área, teniendo en cuenta las necesidades del cliente.

3.1 Implementación de la Base de datos

Posterior al diseño del modelo de datos dimensional se realizó la trasformación al modelo físico, el cual facilita la descripción del almacenamiento de los datos y la relación existente entre las tablas que lo componen.

3.1.1 Estructura de los datos

Una estructura de datos es una vía factible de organizar una colección de datos. La misma se caracteriza por las funciones que se utilizan para acceder y almacenar los elementos individuales de la información. En vísperas de lograr un mejor entendimiento de la solución en el presente trabajo se estructuraron los datos en esquemas que a su vez estarán conformados por tablas.

✓ **Esquemas:** Los esquemas representan una manera de tener organizada toda la información contenida en las BD. Estos pueden contener funciones, operadores y tipos de datos. Esta estructura le permite al usuario tener acceso a ellos siempre y cuando posea los permisos adecuados.

Para el avance de la solución se cuenta con dos esquemas:

- dimensiones: Este esquema presenta las dimensiones comunes propuestas por SIGOB.
- mart_industria_series: En este se encuentran las dimensiones y hechos propios del mercado de datos Series históricas de industria.
- metadatos: En el mismo se encuentran los metadatos de SIGOB.

✓ **Tablas:** Las tablas representan el conjunto de registros que te van a permitir describir un elemento individual de la información.

Para la solución fueron modeladas 10 tablas que se representan en la siguiente tabla que describe dicha estructura:

No.	Esquema	Tablas
1	dimensiones	dim_temporal_anno
2	mart_industria_series	dim_origen
3		dim_destino
4		dim_anno_zafra

5		**dim_indicadores_fund_ind_azucarera**
6		**dim_producciones_industriales_seleccionadas**
7		**hech_indice_vol_fis_origen**
8		**hech_indice_vol_fis_destino**
9		**hech_indicadores_azucareros**
10		**hech_producciones_industriales**
11	**metadatos**	**md_transformation**
12		**md_job**

Tabla 8. Esquemas y tablas

3.1.2 Usuarios y privilegios en la Base de datos

PostgreSQL ofrece la posibilidad de agrupar a los usuarios según las necesidades de permisos y accesos que necesitará cada rol para realizar su función como trabajador del sistema.

✓ **Usuarios y Roles:**

Los usuarios y roles definidos en la BD contribuyen a garantizar la seguridad de la misma. A continuación se describen los roles establecidos:

- **Administrador:** Tiene acceso a la BD en su totalidad, dígase administración y configuración tanto de la BD como de los usuarios restantes.
- **Analista:** Su rol se basa en consultar la información de la BD.
- **Administrador de ETL:** Su función se basa en la realización de los procesos de ETL en la interacción con la BD.

✓ **Privilegios:**

Los privilegios que se les asignan a los usuarios del sistema son basados en el rol que desempeñan. En caso del:

- **Administrador:** A este usuario se le asigna los derechos de *Owner, Select, Update, Insert, Delete, Refresh y Trigger* sobre la estructura de la BD.
- **Analista:** Solo tiene derechos de *Select* los datos almacenados en el Almacén.
- **Administrador de ETL:** A través del rol que este usuario desempeña a la hora de interactuar con la BD, se le asigna el privilegio de *Select, Update, Insert, Delete, Refresh y Trigger,* sobre los datos en el Almacén.

3.2 Implementación del subsistema de integración de datos

En el proceso de integración de datos no es recomendable iniciar el desarrollo de la fuente sin haberla analizado previamente.

La fuente son los ficheros en los que se encuentran almacenados los datos que guardan la información histórica, estos se encuentran en formato .xls y sufrirán un proceso de cambios para facilitar el trabajo con las transformaciones. La etapa de transformación y limpieza es muy importante, una vez que se realiza, la información está lista para ser cargada en la BD. Con la limpieza se detectan los datos erróneos, además de detectar entradas duplicadas y con las transformaciones se combinan y ordenan los datos.

3.2.1 Arquitectura del subsistema de integración

La **figura 20** muestra la arquitectura de integración que se utilizó en el desarrollo de la solución:

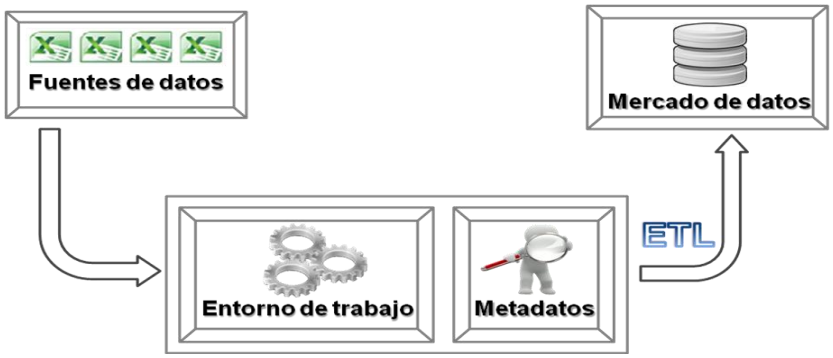


Figura 20. Arquitectura de integración

A continuación se describen los elementos que conforman dicha arquitectura de integración:

- ✓ **Fuente de datos:** Son los datos que se encuentran almacenados en los ficheros fuentes que guardan información histórica de los sistemas.
- ✓ **Entorno de trabajo:** Es donde se preparan los datos para facilitar los procesos de integración y se encuentran todas las transformaciones.
- ✓ **Metadatos:** Es la información que se recoge para controlar como se ejecutan los procesos de integración de datos. Estos datos contienen: campos de la BD de Metadatos
- ✓ **Mercado de datos:** Es el destino hacia donde son cargados los datos después de transformados.

3.2.2 Proceso de Extracción, Transformación y Carga

- ✓ **Extracción de datos**

El primer proceso de ETL consiste en extraer los datos desde los sistemas de origen, que pueden ser provenientes de diferentes fuentes. A través de las fuentes se establece desde dónde se extraerán los datos para analizarlos.

Cada sistema puede usar formatos distintos o una organización diferente de los datos. El formato de la fuente de esta investigación se encuentra en archivos xls. Este proceso convierte los datos a un

formato preparado para iniciar el proceso de transformación.

- ✓ **Transformación de datos**

La transformación es el proceso básico de ETL, se compone de pasos que están enlazados a través de saltos. Los pasos son los elementos más pequeños dentro de las transformaciones. Los saltos son el medio por donde fluye la información entre los diferentes pasos.

Después de realizada la extracción de los datos el sistema se encuentra listo para la etapa de transformación. Durante el proceso se llevaron a cabo tareas tales como: unión por clave, ordenamiento de filas para organizar los campos de las tablas; así como asignación de llaves para relacionar la información de los hechos con las dimensiones.

- ✓ **Carga de datos**

La carga es el último subproceso dentro de los procesos de ETL, el cual consiste en cargar todos los datos que ya han sido transformados satisfactoriamente.

Para una mayor organización en la carga de los datos se realizó primero la carga de las tablas dimensiones que se encuentran en el esquema mart_industria_series, o sea, las que no dependen de la información de otras tablas. Luego se realizó la carga de las tablas hechos al propio esquema que tributan al índice del volumen físico de la industria tanto por el origen como por el destino de los productos, los indicadores fundamentales de la industria azucarera y a las producciones industriales.

En la **figura 21** se muestra la transformación para la carga de la tabla hech_indice_vol_fis_destino, el resto de las transformaciones realizadas podrán ser vistas en los Anexos.

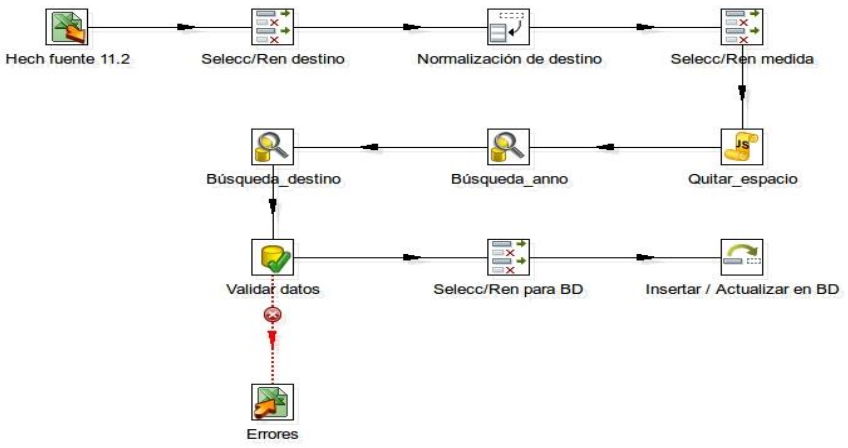


Figura 21. Carga hech_indice_vol_fis_destino

3.2.3 Implementación del Trabajo

Un trabajo es un conjunto sencillo o complejo de tareas con el objetivo de realizar una acción

determinada. En los trabajos se utilizan pasos específicos que son diferentes a los disponibles en las transformaciones. Permite ejecutar una o varias transformaciones de las diseñadas siguiendo una secuencia de ejecución.

Después que se realizaron las transformaciones a los datos se organizó la carga de las tablas estableciendo un orden lógico. Primeramente se cargan las cinco dimensiones propias del mercado, debido a que la tabla dim_temporal_anno es una dimensión compartida, encontrándose cargada en el esquema dimensiones. Seguidamente se cargan los hechos correspondientes al MD. La carga en este orden es de vital importante debido a que se evita cargar llaves nulas que podrían pertenecer a otras tablas que no han sido cargadas.

Seguidamente se muestra en la **figura 22** el trabajo diseñado para el proceso de ETL anteriormente descrito:

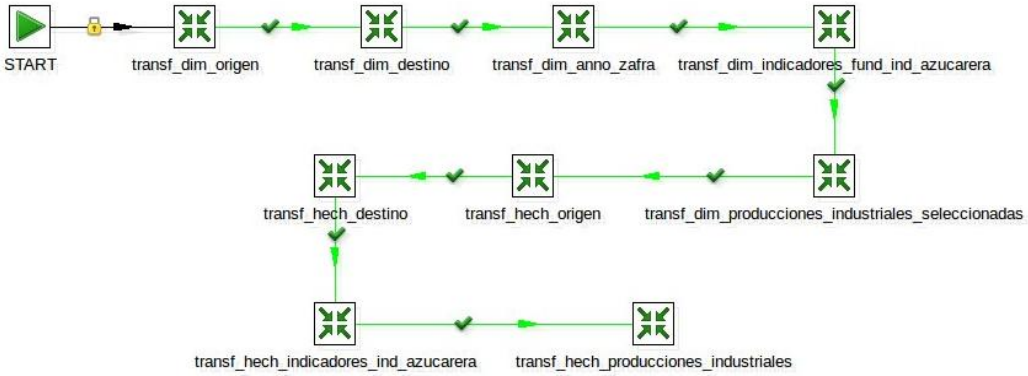


Figura 22. Trabajo del esquema mart_industria_series

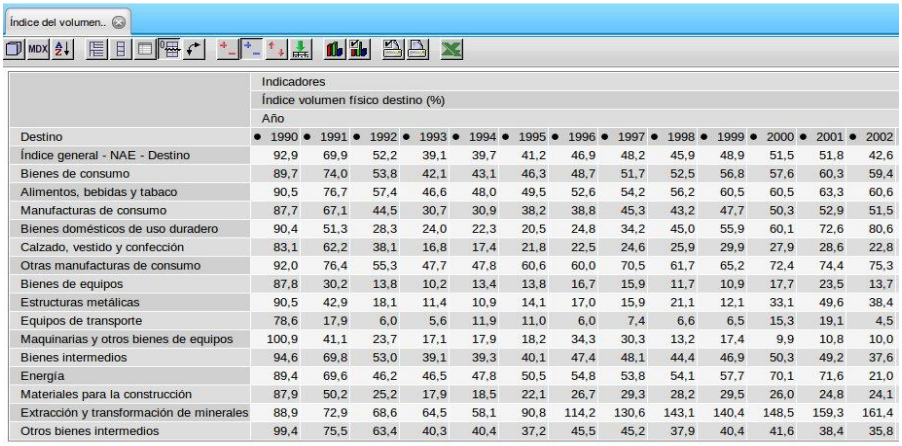
3.3 Implementación del subsistema de visualización

La Inteligencia del Negocio, al igual que otros términos o conceptos no escapa a la diversidad de interpretaciones. Sin embargo queda esencialmente claro que: “no es una metodología, software, sistema o herramienta específica, es más bien una colección de tecnologías que van desde arquitecturas para almacenar datos, metodologías, técnicas para analizar información y software, entre otros, con un fin común para el apoyo a la toma de decisiones”. (24)

3.3.1 Implementación de los reportes candidatos

Las expresiones multidimensionales (MDX) constituyen un lenguaje de consulta para BD multidimensionales, permiten consultar objetos como los 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. Las consultas MDX devuelven un conjunto de celdas que contienen los datos del cubo. (24)

Para la realización de cada reporte se define una consulta MDX, por lo que se tienen 5 consultas, que coinciden con los reportes candidatos. A continuación se muestra en la **figura 23** un ejemplo de cómo queda visualizado el reporte **Índice del volumen físico de la industria por el destino de los productos**, una vez ejecutada la consulta MDX en el Pentaho BI Server.


Índice del volumen..

	Indicadores												
	Índice volumen físico destino (%)												
	Año												
Destino	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Índice general - NAE - Destino	92,9	69,9	52,2	39,1	39,7	41,2	46,9	48,2	45,9	48,9	51,5	51,8	42,6
Bienes de consumo	89,7	74,0	53,8	42,1	43,1	46,3	48,7	51,7	52,5	56,8	57,6	60,3	59,4
Alimentos, bebidas y tabaco	90,5	76,7	57,4	46,6	48,0	49,5	52,6	54,2	56,2	60,5	60,5	63,3	60,6
Manufacturas de consumo	87,7	67,1	44,5	30,7	30,9	38,2	38,8	45,3	43,2	47,7	50,3	52,9	51,5
Bienes domésticos de uso duradero	90,4	51,3	28,3	24,0	22,3	20,5	24,8	34,2	45,0	55,9	60,1	72,6	80,6
Calzado, vestido y confección	83,1	62,2	38,1	16,8	17,4	21,8	22,5	24,6	25,9	29,9	27,9	28,6	22,8
Otras manufacturas de consumo	92,0	76,4	55,3	47,7	47,8	60,6	60,0	70,5	61,7	65,2	72,4	74,4	75,3
Bienes de equipos	87,8	30,2	13,8	10,2	13,4	13,8	16,7	15,9	11,7	10,9	17,7	23,5	13,7
Estructuras metálicas	90,5	42,9	18,1	11,4	10,9	14,1	17,0	15,9	21,1	12,1	33,1	49,6	38,4
Equipos de transporte	78,6	17,9	6,0	5,6	11,9	11,0	6,0	7,4	6,6	6,5	15,3	19,1	4,5
Maquinarias y otros bienes de equipos	100,9	41,1	23,7	17,1	17,9	18,2	34,3	30,3	13,2	17,4	9,9	10,8	10,0
Bienes intermedios	94,6	69,8	53,0	39,1	39,3	40,1	47,4	48,1	44,4	46,9	50,3	49,2	37,6
Energía	89,4	69,6	46,2	46,5	47,8	50,5	54,8	53,8	54,1	57,7	70,1	71,6	21,0
Materiales para la construcción	87,9	50,2	25,2	17,9	18,5	22,1	26,7	29,3	28,2	29,5	26,0	24,8	24,1
Extracción y transformación de minerales	88,9	72,9	68,6	64,5	58,1	90,8	114,2	130,6	143,1	140,4	148,5	159,3	161,4
Otros bienes intermedios	99,4	75,5	63,4	40,3	40,4	37,2	45,5	45,2	37,9	40,4	41,6	38,4	35,8

Figura 23. Reporte del índice volumen físico destino

3.3.2 Navegación de la capa de visualización

Los elementos que componen la estructura de navegación de la información que será exhibida en la capa de visualización del mercado de datos Series históricas de industria, se especifican a continuación, la misma contiene un Área de Análisis (A.A), 4 Libro de Trabajo (L.T) y 5 reportes:

- ✓ **Descripción del Área de Análisis General (A.A.G)**
  - ➢ **A.A.G SIGOB:** Concentra la información de todos los mercados.
- ✓ **Descripción de las Áreas de Análisis (A.A)**
  - ➢ **A.A Series industria:** Contiene la información correspondiente a las Series históricas de industria que son exhibidas, concedidas y aplicadas por los centros históricamente. Posee reportes estadísticos que contribuyen a la concepción de medidas y estrategias encaminadas a realizar un mejor control en el proceso de toma de decisiones en el área antes mencionada.
- ✓ **Descripción de los libros de trabajo**

Seguidamente se les explica cada uno de los Libros de trabajo contenidos dentro del área de análisis Series industria:

- ➢ **L.T Índice volumen físico origen:** Contiene un reporte que permiten realizar un análisis de los datos correspondiente al índice del volumen físico de la industria por el origen de los productos.
- ➢ **L.T Índice volumen físico destino:** Permanece en él un reporte que permiten efectuar un

análisis de la información perteneciente al índice del volumen físico de la industria por el destino de los productos.

- **L.T Indicadores azucareros:** Presenta un reporte que permiten realizar un análisis de los datos correspondiente a los indicadores fundamentales de la industria azucarera.
- **L.T Producciones industriales:** Contiene dos reporte que permiten efectuar un análisis de la información correspondiente a las producciones industriales seleccionadas y la dinámica de dichas producciones.

✓ **Descripción de los reportes**

Consecutivamente se les muestra los reportes contenidos dentro de cada uno de los L.T en el A.A:

- **TS1 -** Índice del volumen físico de la industria por el origen de los productos.
- **TS2 -** Índice del volumen físico de la industria por el destino de los productos.
- **TS3 -** Indicadores fundamentales de la industria azucarera.
- **TS4 -** Producciones industriales seleccionadas.
- **TS5 -** Dinámica de las producciones industriales seleccionadas.

En la **figura 24** se representa el mapa de navegación físico:

Figura 24. Mapa de navegación físico

Conclusiones del capítulo

En el capítulo se realizó un profundo análisis de la fuente de datos y la implementación del MD en general, lo que permitió que se efectuara la carga de los datos a la BD satisfactoriamente, quedando implementado el subsistema de integración de datos. Se modeló el esquema multidimensional, se identificó el área de análisis, los libros de trabajo y los reportes candidatos contenidos. Una vez visualizados cada uno de los reportes van a ayudar a medir el desempeño organizacional de la entidad, alcanzando resultados satisfactorios en el objetivo de la investigación, llegando a conclusiones para la proyección de un mejor trabajo en la organización, que le permita mayores niveles de eficiencia a la ONEI.

Capítulo 4. Validación del mercado de datos

Introducción

En el presente capítulo se realizará la validación de la solución propuesta a través de las distintas vías, para así verificar su correcto funcionamiento y que cuente con la calidad requerida por el cliente. Esta validación se realiza a partir de la aplicación de listas de chequeo y los casos de prueba diseñados, aplicados a la aplicación.

4.1 Calidad del software

La calidad del software es el grado con el que un sistema, componente o proceso cumple los requerimientos especificados y las necesidades o expectativas del cliente o usuario. Es la concordancia del software producido con los requerimientos explícitamente establecidos, con los estándares de desarrollo prefijados y con los requerimientos implícitos no establecidos formalmente, que desea el usuario. (25)

4.1.1 Particularidades que debe presentar el software para que tenga calidad

La calidad del software está determinada por un conjunto de cualidades que lo caracterizan, estas determinan su utilidad y existencia. Para que un producto de software obtenga una buena calidad debe cumplir con las siguientes cualidades:

- ✓ **Mantenibilidad:** capacidad del producto de software de ser modificado. Las modificaciones pueden incluir las correcciones, mejoras o adaptaciones del software a cambios en el ambiente, así como en los requisitos y las especificaciones funcionales.
- ✓ **Funcionalidad:** es la capacidad del software para proporcionar funciones que satisfacen las necesidades declaradas e implícitas cuando el software se usa bajo las condiciones especificadas.
- ✓ **Portabilidad:** capacidad de producto de software de ser transferido de un ambiente a otro.
- ✓ **Confiabilidad:** la capacidad del producto de software para mantener un nivel de ejecución especificado cuando se usa bajo las condiciones especificadas.
- ✓ **Eficiencia:** capacidad del producto de software para proporcionar una ejecución o desempeño apropiado, en relación con la cantidad de recursos utilizados usados, bajo condiciones establecidas.
- ✓ **Usabilidad:** capacidad del producto de software de ser comprendido, aprendido, utilizado y de ser atractivo para el usuario, cuando se utilice bajo las condiciones especificadas. (26)

La Calidad del Software debe aplicarse a medida que el producto vaya siendo construido, paralelo al proceso de desarrollo, se realiza la etapa de pruebas, las cuales constituyen un pilar indispensable para evaluar y determinar la calidad del mismo.

4.2 Prueba de calidad

Las pruebas de software permiten depurar un software desde sus inicios, para que el mismo llegue a manos de los clientes desempeñando todas sus funcionalidades correctamente. Al igual que para el proceso de desarrollo, existen modelos de mejora del proceso de prueba, en la presente investigación se utilizó el Modelo V el cual fue definido por el Centro de Tecnologías de Gestión de Datos (DATEC) con el fin de lograr que el producto posea calidad y así mejorar el proceso de pruebas.

4.2.1 Modelo V

El modelo en V es una variación del modelo en cascada que muestra cómo se relacionan las actividades de prueba con el análisis y el diseño. Como se muestra en la **figura 25**, la codificación forma el vértice de la V, con el análisis y el diseño a la izquierda y las pruebas y el mantenimiento a la derecha.

La unión mediante líneas discontinuas entre las fases de la parte izquierda y las pruebas de la derecha representa una doble información. Por un lado sirve para indicar en qué fase de desarrollo se deben definir las pruebas correspondientes. Por otro sirve para saber a qué fase de desarrollo hay que volver si se encuentran fallos en las pruebas correspondientes.

Por lo tanto el modelo en V hace más explícita parte de las iteraciones y repeticiones de trabajo que están ocultas en el modelo en cascada. Mientras el foco del modelo en cascada se sitúa en los documentos y productos desarrollados, el modelo en V se centra en las actividades y la corrección. (27)

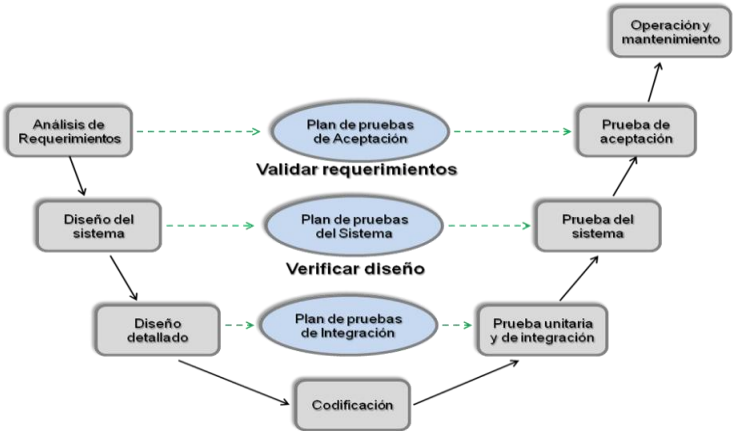


Figura 25. Modelo V

Las pruebas que a continuación se mencionan son aplicadas en el centro DATEC utilizando dicho modelo:

✓ **Prueba unitaria:** Es el proceso de probar los componentes individuales (subprogramas o procedimientos) de un programa. El propósito es descubrir discrepancias entre la especificación de la interfaz de los módulos y su comportamiento real. Estas pruebas son diseñadas y ejecutadas por el desarrollador una vez terminado el desarrollo de cada componente.

✓ **Prueba de integración:** Son las pruebas que se realizan para determinar la integración de los componentes dentro de un sistema y evaluar su correcta interfaz, funcionalidad y desempeño. Estas pruebas son diseñadas y ejecutadas por el desarrollador cuando la solución está completa junto a los especialistas del centro.

✓ **Prueba de sistema:** Son las pruebas que se realizan para determinar el correcto funcionamiento de un sistema y su cumplimiento contra las especificaciones del producto. (27)

✓ **Pruebas de aceptación:** Pruebas que se realizan directamente con el cliente para validar su conformidad con el producto.

Independientemente de estos tipos de prueba que se realizan, el centro efectúa un grupo de pruebas internas encaminadas a la terminación del ciclo de desarrollo del producto las mismas son realizadas por el equipo de desarrollo y especialistas del departamento que participan en la solución.

4.2.2 Casos de prueba

Los casos de prueba verifican si el producto satisface los requerimientos del usuario, tal y como se describe en la especificación de requerimientos y los casos de uso. En la solución propuesta se utilizan varios casos de prueba basados en casos de uso para realizar validaciones concretas. Un ejemplo de estos es el que se muestra en la **figura 26**. (Para más información sobre los mismos ver los artefactos en el expediente de proyecto).

Escenario	Descripción	Variables de Entrada	Variables de Salida	Respuesta del sistema	Flujo central
EC 1. Reporte de prueba del LT-Índice del volumen físico destino	*Muestra el reporte correspondiente al índice del volumen físico por el destino de los productos*	** Año* ** SECTOR/GRUPO*	** Indicadores (Índice volumen físico por el destino de los productos (%))*	*El sistema muestra todas las variables disponibles para el análisis, ubicados en las filas y las columnas que pueden ser visualizadas para cada reporte.*	*Se autentifica.* *En la parte superior izquierda selecciona el AAG. SIGOB.* *Se selecciona el área de análisis de AA.Series industria.* *Se selecciona el libro de trabajo L.T Prueba.* *En la parte inferior izquierda se selecciona Reporte de prueba del LT-Índice del volumen físico destino.* *En el área de trabajo se visualiza el reporte seleccionado.*

Figura 26. Caso de prueba < Presentar el índice del volumen físico destino>

4.2.3 Listas de chequeo

Las listas de chequeo se crean con el fin de concretar y propiciar un buen desarrollo en el trabajo. Son un conjunto de preguntas que sirven para verificar el cumplimiento de los objetivos.

En esta investigación se aplicaron las siguientes listas de chequeo a los artefactos de los procesos

ETL con el fin de evaluar y verificar el potencial de cada uno de ellos, midiendo la confiabilidad y seguridad de los datos cargados:

- ✓ Lista de chequeo del Mapa Lógico de Datos.
- ✓ Lista de chequeo del Diccionario de Datos.
- ✓ Lista de chequeo de Registro de Sistemas Fuentes.
- ✓ Listas de chequeo del Perfilado de Datos

4.2.4 Estructuras de las listas de chequeo

Contienen diferentes indicadores a evaluar, los cuales se encuentran distribuidos en tres secciones:

- ✓ **Estructura del documento**: Contiene todos los aspectos definidos por el expediente del proyecto.
- ✓ **Indicadores definidos por la etapa**: Contiene todos los indicadores a evaluar durante la etapa de análisis de datos.
- ✓ **Semántica del documento**: Contiene todos los indicadores a evaluar respecto a la redacción y ortografía.

A cada uno de los indicadores anteriores los describe los siguientes elementos:

- ✓ **Peso:** Define si el indicador a evaluar es crítico o no.
- ✓ **Indicadores a evaluar:** Son los indicadores a evaluar en las secciones Estructura del documento, semántica del documento e indicadores definidos por las diferentes etapas.
- ✓ **Evaluación:** Es la forma de evaluar el indicador en cuestión. El mismo se evalúa de 1 en caso de que exista alguna dificultad sobre el indicador y 0 en caso de que el indicador revisado no presente problemas.
- ✓ **No procede:** Se usa para especificar que el indicador no es necesario evaluarlo en ese caso.
- ✓ **Cantidad de elementos afectados:** Especifica la cantidad de errores encontrados sobre el mismo indicador.
- ✓ **Comentario:** Especifica los señalamientos o sugerencias que quiera incluir la persona que aplica la lista de chequeo.

Una vez aplicada la lista de chequeo se detectan los indicadores evaluados de mal y con el objetivo de darles solución se especifican en una tabla de No Conformidades (NC), la cual presenta la siguiente estructura:

- ✓ **No:** Es un número consecutivo que indica la cantidad de NC identificadas.
- ✓ **Elemento de evaluación:** Se refiere a un número que identifica al elemento de evaluación para el cual se corresponden los indicadores identificados.
- ✓ **No Conformidad:** Especifica la NC a la que se refiere.

✓ **Fase correspondiente:** Especifica la fase del procedimiento a la que corresponde la NC encontrada.

✓ **Significación:** Especifica si la NC es o no significativa, dependiendo si el indicador es o no crítico.

✓ **Recomendación:** Especifica si la NC es una recomendación, es decir que no es de obligatorio cumplimiento que se solucione por parte de los diseñadores.

✓ **Estado NC:** Especifica el estado de solución en que se encuentra la NC, puede ser Pendiente o Solucionada.

✓ **Respuesta del equipo de desarrollo:** si es necesario se especifica la respuesta que le da el equipo de desarrollo a la NC.

4.2.5 Aplicación de las listas de chequeo

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿El entregable contiene las secciones obligatorias de la plantilla estándar definida para el expediente de proyecto?	0			
crítico	2. ¿El alcance del proyecto describe correctamente los datos de las dimensiones y hechos del MD?	0			
crítico	3. ¿El objetivo expresa correctamente el propósito del documento?	0			
	4. ¿Se hace un uso adecuado del control del documento?	0			
	5. ¿En la sección de acrónimos se definen todos los acrónimos utilizados en el documento?	1		2	Debe especificar el significado de ETL, NAE
	6. ¿En el entregable, la definición de las variables se hace correctamente?	0			

	7. ¿Existe una adecuada correspondencia entre las variables definidas y las descripciones que tienen estas variables?	0			
	8. ¿En el entregable se crea una hoja por cada variable definida?	0			
	9. ¿Queda registrado en el entregable todos los posibles valores que van a tener las variables definidas?	0			
Indicadores definidos en el desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?	0			
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿Se ha identificado errores ortográficos en el entregable?	0			
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	0			
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?	0			

Tabla 9. Aplicación de lista de chequeo al Diccionario de datos

4.3 Resultados de las pruebas

Los resultados obtenidos en las pruebas realizadas a la solución son los siguientes:

✓ **Pruebas unitarias y de integración:** Realizadas por los especialistas del departamento y miembros del equipo de desarrollo arrojaron cinco No conformidades (NC), resueltas todas en el período establecido.

✓ **Pruebas del sistema:** El departamento internamente realizó dichas pruebas arrojando dos NC, ambas resueltas antes de que el producto ingresara a Calisoft. Finalmente el producto fue liberado por Calisoft, lo cual demuestra la calidad del MD Series históricas de industria.

✓ **Pruebas de aceptación:** Inicialmente se detectaron tres NC, una vez resueltas se obtuvo la carta de aceptación del producto por parte del cliente.

Estos resultados se pueden apreciar gráficamente en la **figura 27**.

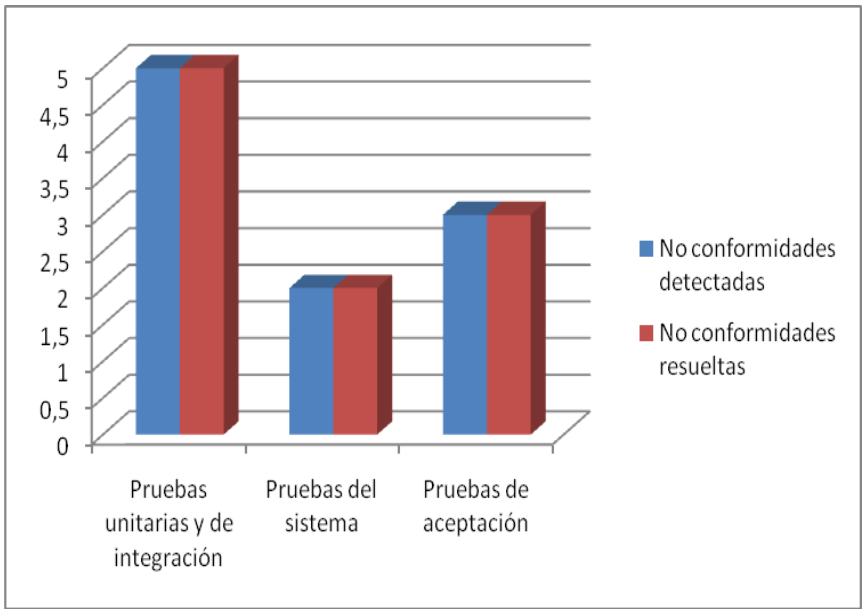


Figura 27. Comportamiento de las pruebas

Conclusiones del capítulo

En el capítulo se realizaron las pruebas necesarias al MD, se utilizaron las herramientas definidas para la aplicación de las pruebas como la lista de chequeo, donde se cubren todos los aspectos aplicables en el tema de análisis y necesarios para cumplir con un buen diseño, además de los casos de prueba garantizando la calidad de la solución y el cumplimiento de los requerimientos del cliente.

Conclusiones generales

Al culminar el presente trabajo de diploma, dándole cumplimiento a los objetivos específicos se arribaron a las siguientes conclusiones:

- En la fundamentación teórica de la investigación se seleccionó la metodología realizada por el Departamento de Almacenes de Datos, la cual guío satisfactoriamente todo el desarrollo del mercado. Además se definieron las herramientas necesarias a utilizar las mismas cumplen con las características pertinentes para el buen progreso del mercado. Tanto la metodología como las herramientas ya han sido utilizadas en el departamento.

- En el análisis y diseño efectuado al MD se identificaron 25 requisitos funcionales, 21 no funcionales y cinco informativos, logrando cumplir con las necesidades del cliente.

- Con la implementación de los subsistemas de almacenamiento, integración y visualización del MD, se obtuvo la disponibilidad de la información del área de industria necesaria para su consulta y análisis por parte de los usuarios, contribuyendo con el proceso de toma de decisiones.

- En la validación de la solución haciendo uso del modelo V, se detectaron 10 no conformidades mediante la utilización de las herramientas de prueba listas de chequeo y casos de prueba encontrándose todas resueltas, obteniendo finalmente la aceptación del cliente.

Recomendaciones

Con la intención de enriquecer la propuesta desarrollada en el presente trabajo de diploma, se exhorta a:

- Desplegar el MD desarrollado en la ONEI.

- Proporcionar cursos de capacitación a los especialistas de la ONEI, para que puedan trabajar con la aplicación satisfactoriamente.

Referencias Bibliográficas

1. **Hernando Velasco, Roberto.** *¨Almacenes de datos (Datawarehouses)¨.* [En línea] [Citado el: 02 de Octubre de 2011.] http://www.rhernando.net/modules/tutorials/doc/bd/dw.html.

2. ¨Qué es un Data warehouse¨. [En línea] [Citado el: 02 de Octubre de 2011.] http://www.gestion.org/documental/gestion-del-conocimiento/4891/que-es-data-warehouse.html.

3. ¨Information management, Datawarehousing datawarehouse y datamart¨. [En línea] [Citado el: 03 de Octubre de 2011.] http://informationmanagement.wordpress.com/2007/10/07/data-warehousing-data-warehouse-y-datamart/.

4. **Inmon, Bill**. *¨El padre del Datawarehousing (DWH)¨.* [En línea] [Citado el: 03 de Octubre de 2011.] http://www.intellego.com.mx/interaccion_articulo.php?idarticulo=47.

5. ¨SQL Max connections, Data Warehousing¨. [En línea] [Citado el: 03 de Octubre de 2011.] http://www.sqlmax.com/dataw1.asp.

6. **Ibarra, María de los A**. *¨Procesamiento Analítico en Línea (OLAP)¨.* [En línea] [Citado el: 2011 de Octubre de 2011.] http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/OLAPMonog.pdf.

7. ¨Almacén de datos¨. [En línea] [Citado el: 05 de Octubre de 2011.] http://issuu.com/jculacio/docs/kddfase1.

8. **Casales Cabrera, María Evelia.** *¨Datawarehouse (Almacenes de datos)¨.* [En línea] [Citado el: 05 de Octubre de 2011.] http://hp.fciencias.unam.mx/~alg/bd/dwh.pdf.

9. **Vega, Liliam; Rojas, Luis y Placeres, Cecilia**. *¨La inteligencia de negocio, Su implantación mediante la plataforma Pentaho¨.* [En línea] [Citado el: 05 de Octubre de 2011.] http://www.google.com.cu/url?sa=t&source=web&cd=7&ved=0CEQQFjAG&url=http%3A%2F%2Fwww.redciencia.info.ve%2Fmemorias%2FProyProsp%2Ftrabajos%2Fl3.doc&rct=j&q=almacen%20de%20datos%20%2B%20aportes&ei=XR2NTt_pD8Lj0QHCpOhY&usg=AFQjCNHVxU5eRTYdu6iEC8nuzQ_c0Tg_wQ&.

10. ¨Procesamiento y modos de almacenamiento de particiones¨. [En línea] [Citado el: 03 de Noviembre de 2011.] http://msdn.microsoft.com/es-us/library/ms174915.aspx.

11. Sinnexus. *¨Datamart¨.* [En línea] [Citado el: 01 de Noviembre de 2011.] http://www.sinnexus.com/business_intelligence/datamart.aspx.

12. ¨Sobre metodología de la ciencia y técnica¨. [En línea] [Citado el: 08 de Noviembre de 2011.] http://www.analitica.com/vam/1999.05/ciencia/03.htm.

13. **González Hernández, Yanisbel.** *¨Propuesta de metodología de desarrollo de Almacenes de datos para DATEC¨.* Cuba : s.n., 2011.

14. ¨Free download manager. Visual Parading for UML¨. [En línea] [Citado el: 07 de Noviembre de 2011.] http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_%28M%C3%8D%29_14720_p/.

15. Tienda Linux. *¨Herramientas gráficas de diseño y administración de bases de datos¨.* [En línea] [Citado el: 08 de Noviembre de 2011.] http://soporte.tiendalinux.com/portal/Portfolio/postgresql_ventajas_html.

16. ¨pgAdmin, Introducción¨. [En línea] [Citado el: 08 de Noviembre de 2011.] http://www.pgadmin.org/.

17. Programming 4 US. *¨PgAdmin III¨.* [En línea] [Citado el: 08 de Noviembre de 2011.] http://mscerts.programming4.us/es/16904.aspx.

18. Gravitar. *¨Características Pentaho BI¨.* [En línea] [Citado el: 07 de Noviembre de 2011.] http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/.

19. ¨Data Cleaner¨. [En línea] [Citado el: 07 de Noviembre de 2011.] http://datacleaner.eobjects.org/.

20. ¨Comparativa BI Open Source¨. [En línea] [Citado el: 08 de Noviembre de 2011.] http://www.stratebi.es/todobi/jun10/Comparativa_OSBI.pdf.

21. Pentaho BI Suite. *¨Pentaho Schema Workbench¨.* [En línea] [Citado el: 08 de Noviembre de 2011.] http://www.google.com.cu/url?sa=t&rct=j&q=pentaho%2Bschema%2Bworkbench%2Bcaracteristicas&source=web&cd=8&ved=0CFUQFjAH&url=http%3A%2F%2Fwww.cognus.cl%2Fmedia%2Fusers%2F1%2F92208%2Ffiles%2F10961%2FPentaho_BI_Suite.pdf&ei=J1S5TtO0F4iA2wWt3PzQBw&usg=AFQjCNHv.

22. ¨Introducción a Apache Tomcat 5.5.¨. [En línea] [Citado el: 08 de Noviembre de 2011.] http://www.google.com.cu/url?sa=t&rct=j&q=apache%2Btomcat%2Bcaracteristicas&source=web&cd=2&ved=0CCUQFjAB&url=http%3A%2F%2Fwww.lsi.us.es%2Fdocencia%2Fget.php%3Fid%3D1923&ei=q1q5TpSuK8PY2AXX_vCfBw&usg=AFQjCNGveTN6jrPIgzaxxKLFRlbYpens8A&cad=rja.

23. ¨Apache Tomcat¨. [En línea] [Citado el: 08 de Noviembre de 2011.] http://www.agapea.com/libros/Tomcat-6-0-La-guia-definitiva-isbn-8441524319-i.htm.

24. ¨Aspectos básicos de las consultas MDX (Analysis Services)¨. [En línea] [Citado el: 04 de Mayo de 2012.] http://msdn.microsoft.com/es-es/library/ms145514.aspx..

25. Computación e Informática. ¨*¿Qué es la calidad del software?¨.* [En línea] [Citado el: 04 de Mayo de 2012.] http://www.rodolfoquispe.org/blog/que-es-la-calidad-de-software.php.

26. ¨Calisoft¨. [En línea] [Citado el: 07 de Mayo de 2012.] http://calisoft.uci.cu/tmp/documentos/normas/iso/NC-ISO-IEC%209126-1.pdf.

27. ¨Ciclo de vida-Modelo V¨. [En línea] [Citado el: 08 de Mayo de 2012.] http://spanishpmo.com/index.php/ciclos-de-vida-modelo-en-v/.

Bibliografía

♦ **Cespedes, Dianelis y González, María L.** ¨Mercado de datos Ciencia e innovación tecnológica¨. [En línea] [Citado el: 02 de Noviembre de 2011.] https://repositorio.datec.prod.uci.cu/svn/cbd_almacenes/Proyectos/SIGOB%20-%20DWH_Estadistico/Datamart_Ciencia_Innovación/.

♦ **Tamargo, L.L.C.** Diseño Físico del Data Warehouse, La revista del empresario cubano. 2011.

♦ ECURED. [En línea] [Citado el: 10 de Mayo de 2012.] http://www.ecured.cu/index.php/Almac%C3%A9n_de_Datos.

♦ **Ferreira, Jose Schmidt y Keyla.** Sistema de Informació. Pentaho. 2009.

♦ Free Download Manager. Free Download Manager. [En línea] [Citado el: 10 de Mayo de 2011.] http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_%5Bcuenta_de_Plataforma_de_Java_14715_p/..

♦ **García Caraballo, MSc. Ing. Julio Alfredo.** D.I.J.A.M.M. El Proceso de Inteligencia Empresarial en las Empresas del Grupo de Diseño e Ingeniería de la Construcción. REVISTA DE ARQUITECTURA E INGENIERÍA. 2010, Vol. IV..

♦ PostgreSQL Global Development Group. [En línea] [Citado el: 17 de Noviembre de 2010.] http://www.postgresql.org..

♦ Portada sobre la plataforma Pentaho Open Source Business Intelligence, El editor de las sentencias MDX (MDX Query Editor). [En línea] [Citado el: 10 de Noviembre de 2011.] http://pentaho.almacen-datos.com/jpivot-editor-mdx.html..

♦ Portada sobre la plataforma Pentaho Open Source Business Intelligence, La plataforma Pentaho Open Source Business Intelligence. [En línea] [Citado el: 2011 de Octubre de 25.] http://pentaho.almacen-datos.com/..

♦ **Salazar, Ricardo Luján.** Datawarehouse para la prestación del servicio público de información estadísticas. México : s.n.

♦ Sinnexus, Bases de datos OLTP y OLAP. [En línea] [Citado el: 10 de Octubre de 2011.] http://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx.

♦ Sinnexus, Datamart. [En línea] [Citado el: 20 de Noviembre de 2011.] http://www.sinnexus.com/business_intelligence/datamart.aspx.

♦ Sinnexus, Datawarehouse. [En línea] [Citado el: 02 de Octubre de 2011.] http://www.sinnexus.com/business_intelligence/datawarehouse.aspx.

♦ Sinnexus, ¿Qué es Business Intelligence? [En línea] [Citado el: 07 de Noviembre de 2011.] http://www.sinnexus.com/business_intelligence/.

Glosario de Términos

AD: Almacén de datos.

DAD: Departamento de Almacenes de datos.

GPL: Licencia Pública General.

MDX: Lenguaje de consulta a estructuras multidimensionales.

UCI: Universidad de las Ciencias Informáticas.

UML: Lenguaje Unificado de Modelado.

Base de datos (BD): Conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso.

SQL: Es un lenguaje declarativo de acceso a BD relacionales que permite especificar diversos tipos de operaciones en éstas.

Indicadores: Magnitud utilizada para medir o comparar los resultados efectivamente obtenidos, en la ejecución de un proyecto, programa o actividad. http://www.definicion.org/indicador

Estadísticas: Ciencia que estudia la recolección, análisis e interpretación de datos.