Universidad de las Ciencias Informáticas

Facultad 6

***Título*:** Mercado de datos Distribución del área industria manufacturera para el Sistema de Información de Gobierno.

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autores:

Marla Camila Coll González

Reinier González Cruz

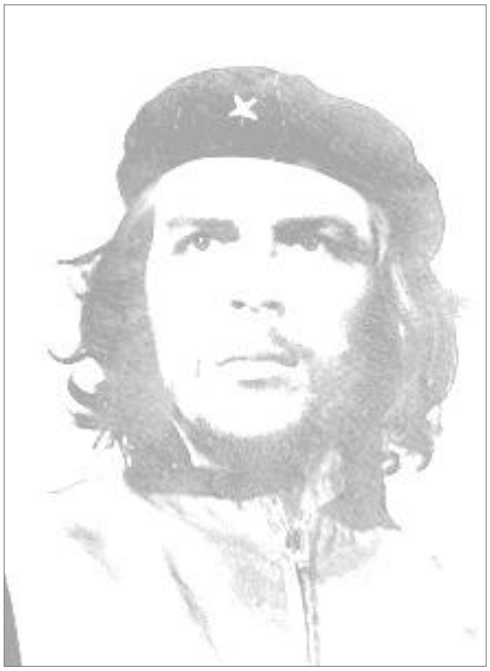
Tutores:

Ing. Yaneysi López Marrero

Ing. Ramón Ernesto Stevenson Borrell

La Habana, Junio 2012

"Año 54 de la Revolución"



No se puede dirigir si no se sabe analizar, y no se puede analizar si no hay datos verídicos; y si no hay todo un sistema de recolección de datos confiables, sin mentiras ni globos, si no hay toda una preparación de un sistema estadístico y de hombres habituados a recoger el dato y transformarlo en números.

Esto es una tarea esencial.

EL CHE

DECLARACIÓN DE AUTORÍA

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

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

_____	_____
Marla Camila Coll González	Reinier González Cruz
Autora	Autor
_____	_____
Ing. Ramón Ernesto Stevenson Borrell	Ing. Yaneisy López Marrero
Tutor	Tutora

Datos de Contacto

Tutor:

Ing. Ramón Ernesto Stevenson Borrell

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Ninguna

Categoría Científica: no

Años de experiencia en el tema: 2

Años de graduado: 3

Correo Electrónico: restevenson@uci.cu

Co-Tutora:

Ing. Yaneisy López Marrero

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Adjunto a la producción

Categoría Científica: no

Años de experiencia en el tema: 2

Años de graduado: 3

Correo Electrónico: yaneisylm@uci.cu

Agradezco:

A mi mamá por haberme brindado su apoyo incondicional, por haberme ayudado en todo momento, por haber confiado en mí pero sobre todo por habérmelo dado todo en la vida.

A mi hermana Anay por haber sido en mi vida más que una hermana, por haber sido mi amiga, mi madre cuando lo necesite y por haberme escuchado siempre.

A mi papá Pedrito por haber sido el único papá que he tenido siempre y por haber sido el mejor del mundo.

A mi cuñado Luis Raciel por todo lo que hizo por mí, por haber tenido paciencia y haberme ayudado siendo lo suficiente crítico para sacar lo mejor de mí, le agradezco que fuera en ocasiones mucho más que mi cuñado, le agradezco por haber sido mi amigo y mi tutor.

A mi tutor Monchy por haber estado junto a nosotros en todo momento, por haber sido nuestra base en estos meses de trabajo siendo más que nuestro tutor, siendo nuestro amigo que nos apoyó y ayudo en momentos difíciles. Gracias profe.

A mi gran amiga de la universidad y que espero sea para toda la vida Diana Rosa, por haber pasado tantos momentos juntos en estos cinco años y por tanto trabajo y ratos felices que pasamos juntas en Artemisa.

A Sonaiti, mi otra amiga de la universidad que ha estado conmigo a lo largo de estos 5 años y que espero que sigamos siendo amigas durante toda la vida.

Marla Camila Coll González

A mi mamá por ser lo más preciado en mi vida, por darme su amor incondicional en todo momento, gracias mami, no estaría escribiendo ahora si no fuese por ti.

AGRADECIMIENTOS

A mi papá por ser mi padre querido, te quiero mucho papá, tengo mucho de ti, gracias por estar siempre que te he necesitado, has estado en los buenos y en los malos, más en los malos que en los buenos y eso vale el doble, gracias papá.

A mi compañera de tesis y compañera en la vida, gracias Marla Camila por todo lo que has hecho por mí, doy gracias por tenerte a mi lado, eres luz en mi vida.

A mi familia por apoyarme siempre, a mis hermanos Rarry y Maide, a mis tíos Mauricio, Emma y Vladimir, a mis primas Baby, Bety, Guelsy y Lilita son todos ustedes el tesoro más preciado.

A mis suegros Pedrito y Georgina que han estado siempre que los he necesitado, en los buenos momentos y en los malos, en especial a ti Georgina que te has portado como una madre para mí.

A mí cuñada Anay por ser tan especial conmigo y a mi concuña Raciel que lo considero como a un hermano, gracias Raciel.

A un gran amigo de la universidad al cual considero como uno de los pocos amigos verdaderos que tengo, por ayudarme cuando lo necesite y cuando no también, por estar en los momentos precisos Yasma, gracias.

A un amigo especial, uno que ha sido el mejor tutor que se podría tener, a ti Monchi por permanecer junto a nosotros en todo momento, por ser más que un tutor, por ser nuestro amigo gracias te doy.

Reinier González Cruz

DEDICATORIA

Le dedico mi trabajo a:

Mi mamá, por haberme apoyado y ayudado durante estos 5 años para que pudiera lograr mi objetivo final, y porque gracias a ella he podido llegar hasta donde estoy.

Mi abuela mima que yo sé que desde donde está me está escuchando y está muy orgullosa de mí,

Mi hermana Anay por darme consejos y por saber levantarme los ánimos en todo momento escuchándome cuando en ocasiones no podía hablar con nadie más.

Mi papá Pedrito por haber estado junto a mí siempre, por haber complacido todas mis malcriadeces, y sobre todo por haber sido mi papá desde los 5 años.

Marla Camila Coll González

Mi mamá, por ser tan buena madre, por darme fuerzas para continuar a pesar de tantas vicisitudes, por existir y porque gracias a ti mami soy la persona que soy, con mis virtudes y defectos.

Tres personas que no están presentes hoy en mi vida pero que lo estuvieron una vez, a estas tres personas que me dieron desde niño todo el amor y el cariño que se le puede dar a alguien, a estas tres personas que me enseñaron a querer, a reír y sobre todo a amar, a estas tres personas que representan mi alma, mi espíritu y mi verdad, con todo el amor que me dieron una vez les dedico mi trabajo a mi tía Dora Hilda mi PANCHA querida, a mi abuelo MAYÉ y a mi abuela AURORA, si existe ese dios en que tanto creían que los acompañe ahora y los cuide por siempre, descansen en paz mis seres queridos, no los voy a olvidar jamás y los voy a querer ETERNAMENTE.

Reinier González Cruz

RESUMEN

Los Almacenes de Datos representan una tecnología avanzada para el manejo de la información y para la generación de reportes. El presente trabajo de diploma se centra en el desarrollo del mercado de datos Distribución del área industria manufacturera en la Oficina Nacional de Estadísticas e Información para el Sistema de Información de Gobierno. En la actualidad la institución cuenta con grandes volúmenes de datos estadísticos no integrados, situación que hace que el trabajo con la información sea arduo y con grandes posibilidades de cometer errores debido a la inconsistencia de los datos a partir de los cuales surge la información estadística. Para dar solución a esta problemática se decide informatizar el proceso de análisis de los datos referentes a la distribución en el área antes mencionada mediante la creación de un mercado de datos que permita el adecuado procesamiento de la información estadística. En la investigación se detallan las metodologías, herramientas y tendencias actuales para el desarrollo de este tipo de soluciones. Además se plantea el análisis, diseño e implementación del antes mencionado mercado de datos.

Palabras claves: Almacén de datos, Mercado de datos.

TABLA DE CONTENIDO

TABLA DE CONTENIDO

Introducción

La información es considerada un recurso estratégico de gran valor para el buen desempeño de las organizaciones. Es un elemento del cual se puede extraer conocimiento y satisfacer las necesidades de personas e instituciones, razón por la cual adquiere una importancia significativa para el desarrollo, equilibrio y adaptabilidad en cualquier sector del mundo. La habilidad para convertir los datos acumulados durante años en información estratégica e integrada es vital para las instituciones, surgiendo sistemas con nuevas tecnologías que permiten a los usuarios analizar los datos en busca de información que les ayude a tomar decisiones claves.

El control de los datos estadísticos dentro de la infraestructura de un país constituye el eslabón fundamental para la toma de decisiones en los principales sectores socio-económicos. La Oficina Nacional de Estadísticas e Información (ONEI) es el órgano rector de la estadística en Cuba y la responsable de gestionar los principales indicadores de la actividad socio-económica. Una de las áreas comprendidas en el análisis estadístico corresponde a la industria manufacturera. Este sector de la industria es el encargado de la actividad económica que transforma una gran diversidad de materias primas en diferentes artículos para la distribución y el consumo. Dentro de la industria manufacturera se encuentran inmersos los procesos de distribución de los productos que ahí se generan. La distribución de los productos se divide en dos esferas: minorista (se encarga de vender los productos al consumidor final) y mayorista (distribuye a empresas o distribuidor minorista pero nunca al comprador concluyente).

La ONEI en la actualidad se enfrenta a una serie de problemas con respecto a la gestión de los datos. A continuación se enumeran algunas de las deficiencias que se generan en los procesos de distribución del área antes mencionadas:

- Utilización de herramientas de recolección de información basada en aplicaciones informáticas de oficina que generan un cúmulo de documentos digitales.
- Los datos posteriores al año 2011 se encuentran almacenados en una base de datos de difícil acceso y comprensión de los datos.
- Información no integrada debido a las disímiles fuentes de datos.
- Los datos no pueden ser consultados a no ser por un especialista de la informática y de la información con alto dominio del negocio.
- Se generan ficheros anuales ocasionando dificultades en la obtención de información estadística a partir de los datos contenidos en dichos ficheros.
- Proceso de recuperación y elaboración de informes costoso en esfuerzo y tiempo.

Todas estas deficiencias provocan que el análisis de los datos sea muy complejo influyendo negativamente en el análisis estadístico de las distintas variables que abarcan los procesos de distribución en el área de la industria manufacturera, aumentando las probabilidades de ocurrencia de errores estadísticos de la información generada. La búsqueda de mejoras en las formas de almacenar, recuperar y presentar la información proveniente de los organismos, tales como principales reportes, cruces de variables, indicadores, porcentajes y demás aspectos de interés es una necesidad urgente para aumentar la disponibilidad de información y mejorar el proceso de toma de decisiones en el área antes mencionada.

La información estadística que maneja la ONEI representa una fuente de incalculable valor para el país, por lo que se hace necesario la integridad y limpieza de los datos, lograr que el acceso a estos se haga de forma rápida y sencilla para así evitar valores erróneos en la información que podrían afectar gravemente la toma de decisiones.

Debido a estas necesidades se plantea la siguiente interrogante como **problema de la investigación**: ¿Cómo contribuir a la toma de decisiones en el proceso de Distribución del área industria manufacturera para el Sistema de Información de Gobierno?

Por lo antes mencionado se define como **objeto de estudio** los almacenes de datos enmarcado en el **campo de acción** mercado de datos Distribución del área industria manufacturera para el Sistema de Información de Gobierno.

La investigación tiene como **objetivo general** desarrollar el mercado de datos Distribución del área industria manufacturera para el Sistema de Información de Gobierno, que contribuya a la toma de decisiones.

Objetivos específicos

- Fundamentar la selección de la metodología, herramientas y tecnologías a utilizar para el desarrollo de los almacenes de datos.
- Realizar el análisis y diseño del mercado de datos distribución del área industria manufacturera.
- Implementar el mercado de datos distribución del área industria manufacturera.
- Validar el mercado de datos distribución del área industria manufacturera.

Tareas de la investigación

- Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de los almacenes de datos.
- Levantamiento de requisitos.
- Descripción de los casos de uso del mercado de datos.
- Definición de los hechos, las medidas y las dimensiones del mercado de datos.
- Diseño del modelo de datos.
- Definición de la arquitectura del mercado de datos.
- Diseño del subsistema de integración.
- Diseño del subsistema de visualización.
- Diseño de los casos de prueba.
- Implementación del modelo de datos.
- Implementación del subsistema de integración.
- Implementación del subsistema de visualización.
- Aplicación de las listas de chequeo.
- Aplicación de los casos de prueba.

Posibles resultados:

- Mercado de datos poblado que contenga toda la información referente a los procesos de distribución que se realizan en el área de la industria manufacturera.
- Capa de visualización de los datos que permita a los usuarios finales interactuar de forma fácil y dinámica con la información.

Estructura del documento

Capítulo 1. Fundamentación teórica de los almacenes de datos: aborda lo referente al estado del arte, actualidad del tema de la investigación, principales tendencias. Presenta una caracterización de la metodología, herramientas y tecnologías a utilizar en el desarrollo de un AD.

Capítulo 2. Análisis y diseño del mercado de datos Distribución del área industria manufacturera: contiene lo referente a los requisitos, reglas del negocio, modelo de datos y casos de uso. Además se realiza el diseño de los subsistemas de almacenamiento, integración y visualización.

Capítulo 3. Implementación del mercado de datos Distribución del área industria manufacturera: aborda lo relacionado con los procesos de almacenamiento, así como los procesos de Extracción, Transformación y Carga (ETL) y de Inteligencia de Negocio (BI por sus siglas en inglés *Bussines Inteligence*).

Capítulo 4. Validación del mercado de datos Distribución del área industria manufacturera: trata acerca de la validación de la solución mediante las listas de chequeo y los casos de prueba.

Capítulo 1. Fundamentación teórica

1.1 Estado del arte

Las bases de datos constituyen en la actualidad una herramienta imprescindible en cualquier ámbito social, económico o científico. Con el avance de la tecnología y el incremento de la información esta herramienta pasó a jugar un valioso papel en el desarrollo de la mayoría de las empresas e instituciones. A medida que los volúmenes de datos crecían, las instituciones comenzaron a almacenar los datos en diferentes fuentes de información, lo que trajo consigo que el análisis y el acceso a estos fuera cada vez más engorroso; de ahí surge la necesidad de integrar la información en un solo destino. Los almacenes de datos surgen como solución a esta problemática y son una colección de datos integrados y de fácil acceso, por lo que constituyen un apoyo fundamental para la toma de decisiones.

1.1.1 Antecedentes históricos y actualidad

Se distinguen tres etapas fundamentales en la historia de los almacenes de datos:

Principio de los ochenta

La tecnología de esos tiempos trataba de automatizar los procesos repetitivos o administrativos utilizando los sistemas OLTP (Procesamiento de transacciones en línea). Estos sistemas de información solo resolvían los problemas de carácter operativo. El tipo de decisiones que soportaban estaban asociadas con la ejecución y el apoyo de las tareas básicas del negocio, por lo tanto, no eran muy útiles para realizar análisis avanzados de tipo estratégico (1).

De mediados a finales de los ochenta

En esta etapa los requerimientos de información en las empresas fueron cambiando ya que los responsables de los diferentes departamentos precisaban más informes. La sección informática de las instituciones no respondía a las necesidades de información que se les exigía. De ahí surgió la necesidad de almacenar toda la información en una base de datos detallada y con ello surgieron nuevos inconvenientes:

- ¿Cómo cargar los datos de las diferentes fuentes e integrarlos?
- ¿Cómo acceder a la información de forma ágil y eficiente?

En el año 1985 aparece el primer AD realizado por Bill Inmon (científico de la computación, reconocido por muchos como el padre de los almacenes de datos).

Desde finales de los 90 hasta la actualidad

En los últimos años los almacenes de datos han ido perfeccionándose según las necesidades de los clientes, de manera tal que en la actualidad existen AD que proporcionan soluciones satisfactorias para todo tipo de usuarios.

1.1.2 Actualidad cubana

La inteligencia de negocios en Cuba, término que se encuentra inmerso en los almacenes de datos, juega un papel importante puesto que se ha ido extendiendo a un número mayor de instituciones a través de los años. Los primeros avances fueron en el ámbito comercial, ya que las empresas de este sector cuentan con el presupuesto y la visión estratégica necesaria para asumir las primeras experiencias. Seguidamente se han incorporado otras entidades de disímiles sectores, entre ellos se destaca la rama de las comunicaciones, por la calidad y los niveles de desarrollo asociados a sus proyectos. La tendencia que se ha seguido en cuanto a los sistemas de BI, es la de los almacenes de datos, teniendo en cuenta el papel que estos juegan dentro de una solución para el proceso de toma de decisiones (2).

Las universidades cubanas representan un eslabón significativo en el avance paulatino de los proyectos de inteligencia de negocio en el país. Específicamente en la capital cubana se destaca la Universidad de la Habana, seguida del Instituto Superior Politécnico “José Antonio Echevarría” y luego la Universidad de las Ciencias Informáticas (UCI). La UCI a pesar de ser una universidad joven, cuenta con un amplio centro de desarrollo y en particular un departamento de almacenes de datos. La universidad debido a las problemáticas que presenta la ONEI firmó un convenio para solucionar los problemas existentes, de donde surge el Sistema de Información de Gobierno (SIGOB). Este sistema está compuesto por un conjunto de soluciones informáticas entre las que se encuentra el AD SIGOB para la ONEI el cual está conformado por una serie de MD, algunos de estos son: el mercado balance económico y financiero, el mercado comercio exterior, el mercado comercio interior, entre otros.

1.2 Fundamentos teóricos de la investigación

Dadas las características de un sistema de AD, su aplicación puede tener variados fines en una diversidad de industrias. Se puede decir que su aplicación más práctica corresponde a entornos de empresas en los que se identifican grandes volúmenes de datos, asociados a: cantidad de clientes, variedad de productos, cantidad de transacciones, entre otros (1).

1.2.1 Definición y características de los almacenes de datos

Existen diversas definiciones de AD: (3)

- Un AD es una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis.
- Un AD provee dos beneficios empresariales reales: integración y acceso a los datos. Los almacenes de datos eliminan una gran cantidad de datos inútiles y no deseados, como también el procesamiento desde el ambiente operacional clásico.
- Un AD es una colección de datos de fácil acceso, alimentado por múltiples fuentes y organizado por temas de información específicos.

Sin embargo el concepto de AD más conocido es el propuesto por Bill Inmon (1996) que define los almacenes de datos con cuatro características fundamentales: (3)

Temáticos: los almacenes de datos al estar orientados hacia un determinado ámbito se organizan por temas específicos, a diferencia de los sistemas relacionales que se organizan por procesos funcionales, nóminas, entre otros. Esta forma de ordenar los datos permite a los usuarios disponer de la información específica que se necesite en un momento determinado.

Integrados: se trata de una base de datos única que contiene toda la información proveniente de las distintas fuentes integradas en un solo destino. Esta característica permite el acceso a la información de manera más sencilla y garantiza un ahorro de esfuerzo y tiempo en el análisis de los datos.

No volátiles: los almacenes de datos al tener como objetivo fundamental el apoyo a la toma de decisiones poseen como principal característica que sus datos nunca son borrados ni actualizados. Las operaciones que se realizan en un AD son las de carga y consulta.

Variables en el tiempo: los datos al no ser actualizados, son almacenados históricamente, lo que permite obtener el conjunto de valores que estos han tenido a lo largo del tiempo, con el fin de identificar y analizar el valor que ha ido tomando el dato en una etapa determinada.

1.2.2 Mercado de datos

Un MD se define como una base de datos departamental creada para un área específica de una entidad y es alimentado por la integración de un conjunto de fuentes de información. Para poder analizar satisfactoriamente la información generada por un determinado sector de una empresa es necesario encontrar la manera más adecuada de estructurar sus datos. Esta estructura puede estar montada sobre una base de datos OLTP o una base de datos OLAP (por sus siglas en inglés *On-Line Analytical Processing*).

Base de datos OLTP

Los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones. Una transacción genera un proceso atómico (que debe ser validado o invalidado), y que puede involucrar operaciones de inserción, modificación y borrado de datos. El proceso transaccional es típico de las bases de datos operacionales. A continuación se presentan algunas de las características de estos sistemas: (4).

- El acceso a los datos está optimizado para tareas frecuentes de lectura y escritura. (por ejemplo: la enorme cantidad de transacciones que tienen que soportar las bases de datos de bancos o hipermercados diariamente).
- Los datos se estructuran según el nivel aplicación.
- Los formatos de los datos no son necesariamente uniformes en los diferentes departamentos (es común la falta de compatibilidad).
- El historial de datos suele limitarse a actuales o recientes.

Base de datos OLAP

Se basan en los cubos OLAP, los cuales 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. Los MD diseñados a partir de bases de datos OLAP tienen como principales características: (4)

- El acceso a los datos suele ser de solo lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o borrado.
- Los datos se estructuran según las áreas de negocio, y los formatos de los datos están integrados de manera uniforme en toda la organización.
- El historial de datos es a largo plazo.
- Las bases de datos OLAP se suelen alimentar de información procedente de los sistemas operacionales existentes mediante los procesos de ETL.

Clasificación de los sistemas OLAP

Procesamiento Analítico Multidimensional (MOLAP)

La arquitectura MOLAP usa bases de datos multidimensionales para proporcionar el análisis, su principal premisa es que OLAP está mejor implantado almacenando los datos multidimensionalmente. El sistema MOLAP utiliza una arquitectura de dos niveles: la base de datos multidimensional y el motor analítico. La base de datos multidimensional es la encargada del manejo, acceso y obtención del dato (4).

Procesamiento Analítico Relacional (ROLAP)

La arquitectura ROLAP accede a los datos almacenados en un AD para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales. El sistema ROLAP utiliza una arquitectura de tres niveles. La base de datos relacional maneja los requerimientos de almacenamiento de datos, y el motor ROLAP proporciona la funcionalidad analítica. El nivel de base de datos usa bases de datos relacionales para el manejo, acceso y obtención del dato. El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios (4).

Procesamiento Analítico Híbrido (HOLAP)

Un desarrollo un poco más reciente ha sido la solución OLAP híbrida, la cual combina las arquitecturas ROLAP y MOLAP para brindar una solución con las mejores características de ambas: desempeño superior y gran escalabilidad. HOLAP permite almacenar una parte de los datos en un sistema MOLAP y el resto como en uno ROLAP (4).

Para el desarrollo del MD Distribución del área industria manufacturera se utilizará el modo de almacenamiento ROLAP debido a las características que presentadas y además porque el gestor de base de datos PostgreSQL, definido para el desarrollo de la solución no soporta bases de datos multidimensionales.

1.2.3 Ventajas y desventajas de los almacenes de datos

Los almacenes de datos aportan importantes beneficios ya que su desarrollo propicia que la información sea accesible, correcta, uniforme e integrada. Todas estas características brindan una serie de ventajas: (5)

- Menor consumo de tiempo al ser las consultas más sencillas y precisas que si se consultaran bases de datos relacionales.
- Generación dinámica de consultas ya que no es necesario anticipar las consultas que se van a realizar.

- Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.
- Incremento de las capacidades para atender las necesidades de los clientes.
- Resuelve los problemas de integridad y calidad de datos.
- Permite que los usuarios accedan a la información en línea, contribuyendo a la efectividad para operar en las tareas.

A pesar de las ventajas de los almacenes de datos, estos también traen consigo una serie de inconvenientes: (5)

- Su elaboración resulta muy costosa debido a los recursos que se emplean para su desarrollo.
- La creación de los almacenes suele llevar tiempo debido a la cantidad de información que se necesita recoger para que cumpla su objetivo.
- Aumentar el número de dimensiones supone incrementar exponencialmente el tamaño de la base de datos.

Tabla 1. Diferencias entre las bases de datos relacionales y los almacenes de datos (1).

Base de datos relacional	Almacén de datos
Predomina la actualización.	Predomina la consulta.
La actividad más importante es de tipo operativo (día a día).	La actividad más importante es el análisis y la decisión estratégica.
Predomina el proceso puntual.	Predomina el proceso masivo.
Mayor importancia a la estabilidad.	Mayor importancia al dinamismo.
Datos en general desagregados.	Datos en distintos niveles de detalles y agregación.
Importancia del dato actual.	Importancia del dato dinámico.
Importancia del tiempo de respuesta de la transacción instantánea.	Importancia de la respuesta masiva.
Estructura relacional.	Visión multidimensional.
Usuarios de perfiles medios o bajos.	Usuarios de perfiles altos.
Explotación de la información relacionada con la operativa de cada aplicación.	Explotación de toda información externa o interna de todo el negocio.

1.2.4 Etapas de un AD

Para desarrollar satisfactoriamente un AD es necesario transitar por tres etapas fundamentales que se describen a continuación: análisis y diseño, ETL y por último la etapa de inteligencia de negocio.

Etapa de análisis y diseño

Esta etapa se centra en el análisis del negocio, especificando los requerimientos que debe cumplir el sistema. Debe quedar plasmado el modelo de datos dimensional con los respectivos hechos, dimensiones y medidas que conformarán el MD.

Los **hechos** son las variables de negocio sobre los que se va a totalizar, promediar, y en general realizar operaciones de agregación que conduzcan a conclusiones sobre la evolución del área o departamento que se estudie. La denominación estadística de dichas variables de negocio sería la de variables cuantitativas (6).

Una **dimensión** es una característica que hace referencia a un hecho y que aporta información relevante para un análisis posterior que contribuya a la toma de decisiones (6).

Las **medidas** son valores calculables o físicos que aportan información acerca de un comportamiento específico en el negocio. Son estas las que permiten al usuario conocer acerca de las estadísticas o tendencias en un determinado ámbito y que aportan información relevante (6).

Las **tablas de hechos** contienen medidas y dimensiones que aportan información permitiendo la comprensión del negocio. También se encuentran presentes las medidas calculables que arrojan valores cuantitativos ayudando al proceso de toma de decisiones.

Esquemas de los almacenes de datos

Durante la etapa de análisis y diseño debe quedar definido el esquema a emplear en la realización del MD. Los esquemas de los almacenes de datos se refieren a la forma de estructurar la información. En la arquitectura de un AD se definen tres esquemas:

Esquema estrella: es el esquema más sencillo en la arquitectura de un AD. En este diseño la tabla hecho está rodeada por todas las dimensiones formando una estructura, lo que hace más simple la implementación de mecanismos básicos para poder utilizarla como una herramienta de consultas OLAP. El motivo de dejar de mantener las tablas en el modelo relacional y permitir el almacenamiento de información redundante es optimizar el tiempo de respuesta de las bases de datos y dar información a un usuario en menor tiempo posible. En este modelo, para obtener información solicitada no hay que construir una sentencia SQL (lenguaje de consulta estructurado por sus siglas en inglés *structured query language*) muy compleja que lea muchas tablas de una vez. Una herramienta de consultas solo tiene que acceder a una tabla (7).

Esquema copo de nieve: el objetivo de este esquema está orientado a facilitar el mantenimiento de las dimensiones. El uso más común de esta arquitectura es cuando las tablas de dimensiones son muy grandes o complejas y es muy difícil representar los datos en esquema estrella. Para extraer los datos de las tablas en esquema copo de nieve hay que vincular las tablas en las sentencias SQL (7).

Esquema constelación: para cada esquema estrella o esquema copo de nieve de un AD es posible construir un esquema constelación de hechos. Este esquema es más complejo que las otras arquitecturas debido a que contiene múltiples tablas de hechos. Con esta solución, las tablas de dimensiones pueden estar compartidas entre más de una tabla de hechos. El esquema constelación de hechos tiene mucha flexibilidad y esa es su gran ventaja (7).

Etapa de ETL

ETL es el proceso que organiza el flujo de los datos entre los diferentes sistemas en una organización y aporta los métodos y herramientas para mover datos desde múltiples fuentes a un AD, limpiarlos y cargarlos en otra base de datos o MD. A continuación se describen las fases del proceso de ETL:

Extracción: la primera fase en el proceso de ETL consiste en extraer los datos de las diferentes fuentes en que se encuentran y dejarlos listos para el proceso de trasformación. Habitualmente en una organización los datos se encuentran almacenados en bases de datos relacionales o ficheros planos, aunque también se pueden recopilar en otras estructuras diferentes.

Transformación: generalmente las instituciones no cuentan con aplicaciones únicas para el almacenamiento de los datos sino que pueden tener distintos sistemas para atender un mismo conjunto de operaciones y en esos casos es probable que las fuentes contengan datos duplicados, a veces erróneos, redundantes o incompletos. En la fase de transformación se llevan a cabo los procesos de limpieza, estandarización e integración de los datos con el fin de convertirlos en información apta para ser cargada y posteriormente analizada.

Carga: la fase de carga es el momento en que los datos provenientes de la fase de transformación son incluidos en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes. Al realizar esta operación se aplicarán todas las restricciones que se hayan definido, por ejemplo: valores únicos, integridad referencial, campos obligatorios, rangos de valores, entre otros. Estas restricciones están bien definidas, en tanto garantizan la calidad de los datos en el proceso de ETL (8).

Etapa de inteligencia de negocio

La inteligencia de negocio es el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos e información desestructurada en información estructurada para su explotación directa o para su análisis. La inteligencia de negocio actúa como un factor estratégico para una empresa u organización, generando una potencial ventaja competitiva, que no es otra que proporcionar información privilegiada para responder a los problemas de negocio. El objetivo fundamental de esta etapa es propiciar el análisis de la información con el fin de favorecer a mejores resultados en la toma de decisiones, para ello se crean los cubos OLAP, a través de los cuales es posible realizar los reportes y consultas que permiten la comprensión de los datos (4).

1.2.5 Metodología

En la actualidad existen disímiles metodologías a seguir en la construcción de un AD, pero las más utilizadas son las metodologías de Bill Inmon y Ralph Kimball. La mayor diferencia entre los dos autores es el sentido de la construcción del AD, esto es comenzando por los MD o ascendente (*Bottom-up*) según Kimball o comenzando con todo el AD desde el principio o descendente (*Top-Down*) según Inmon. La metodología de Inmon se basa en conceptos bien conocidos del diseño de bases de datos relacionales. La metodología para la construcción de un sistema de este tipo es la acostumbrada para construir un sistema de información utilizando las herramientas habituales, al contrario de la de Kimball, que se basa en un modelado dimensional (9).

Para la creación del MD se utilizará la propuesta de metodología para el desarrollo de AD en el centro de Tecnologías de Gestión de Datos (DATEC), que toma como base la metodología de Ralph Kimball debido a las grandes ventajas que representa para la creación de una solución de AD efectiva y con la calidad requerida. Esta metodología robusta crea los conceptos de hechos y dimensiones lo que aporta un apoyo fundamental al proceso de toma de decisiones. El hecho de ser una metodología ascendente representa una gran ventaja dado que coincide con la división lógica de las empresas, instituciones, entre otros. Es una metodología madura y reconocida, además de existir abundante documentación sobre la misma.

A pesar de todas las ventajas que ofrece la utilización de la metodología de Ralph Kimball, esta no es totalmente adaptable a las características del centro y de la producción en la UCI, por lo que se decidió utilizarla como guía en el proceso de confección de metodología de desarrollo del departamento de AD. Entre las principales desventajas de la metodología de Ralph Kimball se encuentran: (10)

- No tiene definido un criterio que permita estimar los costos de desarrollo de un AD, basándose en las características de la construcción del mismo.

- Presenta un grupo de roles, pero no explica claramente cuáles son las competencias y responsabilidades de cada uno dentro del proyecto. Por la cantidad de roles que propone se necesita de grupos grandes para su desarrollo.
- Propone un gran número de actividades y artefactos que pueden extender los tiempos de desarrollo si se cuenta con pocos recursos humanos, además no se especifica cómo deben realizarse estos artefactos.
- Está estructurada para el desarrollo de proyectos–productos, donde un proyecto desarrolla un producto determinado.
- No establece el análisis de diferentes criterios de diseño en el levantamiento de requisitos que permita la construcción más adecuada del almacén, teniendo en cuenta las metas de la organización, las necesidades de los usuarios y la disponibilidad de las fuentes operacionales.

Por tales motivos se decidió definir una metodología que permita mitigar las desventajas anteriormente expuestas y que se ajuste a las condiciones y características de producción de DATEC y de la UCI. Complementando la metodología de Kimball se hace referencia a lo planteado en la tesis de doctorado de Leopoldo Zenaido Zepeda Sánchez, orientando así el trabajo hacia los casos de uso y estando más guiados hacia las tendencias y normas del centro (10).

La elección de un ciclo de vida adecuado para cada desarrollo está relacionada con las características del producto a obtener a partir de los requisitos y el entorno de desarrollo, como son: aspectos técnicos, características del equipo de desarrollo, tipo de software, condiciones del cliente, entre otras. El ciclo de vida de la metodología está organizado por fases, algunas de ellas pueden ser implementadas de forma paralela según el componente que se está desarrollando para integrarse al final de la solución, esto permite un proceso más ágil y ajustar el desarrollo al modelo que sigue el centro basado en líneas de producto. Cada grupo de trabajo puede desarrollar un componente en específico y después de finalizada su tarea pasar a otros proyectos donde ejecutan la misma función. Además el desarrollo del AD a partir de la construcción de MD permite ir avanzando de manera incremental, obteniendo resultados parciales para satisfacer al cliente hasta lograr la solución final (10).

Ciclo de vida de la metodología propuesta

Las fases están definidas teniendo en cuenta las propuestas por la metodología de Ralph Kimball, una serie de procesos, actividades y características de desarrollo de los proyectos de software en la UCI. Se definieron ocho fases que se describen a continuación: (10)

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, con el fin de determinar qué es lo que se desea construir y qué condiciones existen para el desarrollo de la misma.

Requerimientos: se identifican las necesidades de información y reglas del negocio; haciendo un levantamiento detallado de cada una de las distintas fuentes de datos a integrar. Se definen los requerimientos a partir de una comparación de las necesidades y las reglas del negocio con los elementos disponibles en las fuentes.

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, las tecnologías a utilizar, hardware y software, entre otros aspectos de gran importancia.

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 para obtener los reportes, gráficos, mapas y otros que cubran los requerimientos firmados con el cliente final.

Prueba: se realizan varias pruebas, comenzando por las pruebas de unidad, luego las pruebas de integración y sistema, hasta las pruebas de aceptación con el cliente final.

Despliegue: en este flujo se configuran los servidores necesarios y se instalan las herramientas según la arquitectura definida y se realiza la carga histórica de los datos.

Soporte y Mantenimiento: se brindan los servicios de soporte y mantenimiento a los usuarios finales.

Gestión y Administración del Proyecto: este flujo transcurre a lo largo de todo el ciclo de vida, aquí es donde se controla, gestiona y chequea todo el desarrollo.

1.2.6 Herramientas

Herramientas de modelado

Visual Paradigm 6.4

Visual Paradigm para el Lenguaje Unificado de Modelado (UML) es una herramienta profesional que soporta el ciclo de vida completo del desarrollo de software, análisis y diseño orientados a objetos, construcción, pruebas y despliegue. El software de modelado UML ayuda a una rápida construcción de

aplicaciones de calidad. Permite dibujar todos los tipos de diagramas de clases, código inverso, generar código desde diagramas y generar documentación. Las principales características de la herramienta son: (11)

- Soporta aplicaciones web.
- Varios idiomas.
- Generación de código para Java y exportación como HTML (Lenguaje de Marcado de Hipertexto).
- Fácil de instalar y actualizar.
- Compatibilidad entre ediciones.

Gestor de base de datos

PostgreSQL 9.1

PostgreSQL es un sistema de base de datos relacional, basado en Postgres Versión 4.2, que fue desarrollado en la Universidad de California en Berkeley por el Departamento de Ciencias de la computación. PostgreSQL es un descendiente de código abierto del código original de Berkeley. Es compatible con una gran parte del estándar SQL y ofrece las siguientes características: (12)

- Consultas complejas.
- Claves externas.
- Disparadores.
- Integridad de las transacciones.

Además, PostgreSQL puede ser ampliado por el usuario en muchos aspectos, por ejemplo mediante la adición de nuevos tipos de datos, funciones, operadores, funciones de agregación, métodos de índice y lenguajes de procedimiento (12).

PgAdmin 1.14

Es una de las herramientas más populares para administrar las bases de datos en PostgreSQL que presenta las siguientes características: (13)

- Es un software libre.
- Accede a todos los objetos del PostgreSQL y responde a las necesidades de todos los usuarios, desde escribir consultas SQL simples hasta desarrollar bases de datos complejas.
- Permite que se pueda usar en Linux, FreeBSD, Solaris, Mac OS X y Windows.

Herramientas de ETL

DataCleaner 1.5.3

Es una aplicación de código abierto para el perfilado, validación y comparación de los datos. Es fácil de utilizar, genera sofisticados informes y gráficos que permite a los usuarios comprobar la calidad de los datos existentes, la investigación estadística, la preparación para el proceso de ETL y otras actividades (14).

Pentaho Data Integration 4.2

Herramienta utilizada en la fase de ETL que posee gran utilidad debido a que limpia, integra y carga la información en la base de datos final. Pentaho Data Integration brinda una etapa de ETL poderosa puesto que el uso de Kettle permite evitar grandes cargas de trabajo manual, frecuentemente difícil de mantener y de desplegar. Además de ser de código abierto y sin costos de licencia, las características básicas de esta herramienta son: (14)

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

Herramientas de inteligencia de negocio

Schema Workbench 3.2.0

El esquema de Mondrian Workbench es una interfaz de diseño que permite crear y probar esquemas de cubos OLAP Mondrian visualmente. El motor de Mondrian procesa las solicitudes de MDX (por sus siglas en inglés *Multidimensional Expressions*) con el modo de almacenamiento ROLAP. Estos archivos de esquema XML (por sus siglas en inglés *eXtensible Markup Language,* Lenguaje de marcas extensible) de metadatos son los modelos que se crean en una estructura específica utilizada por el motor de Mondrian. Estos modelos XML se pueden considerar en forma de cubos que utilizan estructuras de hecho existente y tablas de dimensiones que se encuentren en una base de datos (16).

Pentaho BI Server 3.8.0

Es una plataforma que provee el soporte y la infraestructura necesaria para crear soluciones de inteligencia empresarial a problemas de negocios. El marco proporciona los servicios básicos, incluidos autenticación, registro, auditoría, servicios web y motor de reglas. Incluye un motor de solución que

integra reportes, análisis, tableros de comandos y componentes de minería de datos. Funciona como un sistema basado en administración web de informes, el servidor de integración de aplicaciones y un motor de flujo de trabajo ligero. Está diseñado para integrarse fácilmente en cualquier proceso de negocio (17).

Mondrian OLAP Server 3.0.4

Es un servidor OLAP de código abierto que gestiona la comunicación entre la aplicación OLAP y la base de datos. Permite crear cubos de información para análisis multidimensional. Este proporciona la conexión a la base de datos y ejecuta las sentencias SQL (15).

Web Apache Tomcat 5.5

Es un servidor web de aplicaciones que gestiona solicitudes y respuestas HTTP (por sus siglas en inglés *Hypertext Transfer Protocol*, Protocolo de Transferencia de Hipertexto). Es un servidor de aplicaciones o contenedor de Servlets/JSP (por sus siglas en inglés *Java Server Programming*). Es de código abierto, implementado con tecnología Java que le permite funcionar en cualquier sistema operativo que se encuentre la máquina virtual de Java permitiendo a los usuarios realizar reportes (18).

1.3 Conclusiones

A partir del análisis del estado del arte y de la importancia de la utilización de los almacenes de datos para la toma de decisiones se decide implementar una solución que facilite la integridad y disponibilidad de la información y dé solución al problema existente, para ello se obtuvieron los resultados siguientes:

- Se seleccionó la metodología propuesta por el departamento DATEC, siendo esta la adecuada para resolver la problemática por sus ventajas y por estar guiada hacia las normas del centro.
- Se seleccionó para el modelado de datos Visual Paradigm 6.4, como sistema gestor de base de datos PostgresSQL en su versión 9.1 y como aplicación gráfica para la administración de la base de datos PgAdmin 1.14. Para los procesos de BI y ETL se trabajará con las herramientas DataCleaner 1.5.3, Pentaho Data Integration 4.2.0, Schema Workbench 3.2, Pentaho BI Server 3.8, Mondrian OLAP Server 3.0.4 y Web Apache Tomcat 5.5 respectivamente.

Capítulo 2. Análisis y diseño del mercado de datos Distribución del área industria manufacturera

En este capítulo se realiza un análisis profundo del negocio con el objetivo de identificar los requerimientos que debe cumplir la solución. Se identifican las reglas del negocio, los casos de uso y los actores que interactúan con el sistema así como sus principales funcionalidades. Se definen los hechos, dimensiones y medidas por los que quedará conformado el modelo de datos. Se diseñan los subsistemas de integración y visualización de datos de la aplicación.

2.1 Caracterización de las áreas de la organización

La ONEI se encarga de manejar los indicadores estadísticos del país ejerciendo una adecuada dirección, ejecución y control de la captación de cifras económicas y sociales. La información estadística se recoge a través de centros informantes por municipios y provincias y se almacena en archivos con extensión dbf que son generados con una determinada periodicidad (mensual, trimestral, semestral, anual). Esta información es registrada en modelos organizados por organismos y sus respectivos centros informantes. Los modelos están compuestos por un conjunto de indicadores ajustados a cada centro informante en dependencia del trabajo que se realice en cada uno. Los modelos 0729 y 0006 contienen la información referente a los indicadores seleccionados de la variante de industria.

El **modelo 0006** tiene como objetivo captar una selección de indicadores específicos que caracterizan el comportamiento de la industria, tanto en la producción de productos terminados como de materias primas. Esta información se emplea en la elaboración del Informe de la Economía Nacional, el Anuario Estadístico de Cuba y en otros trabajos estadísticos. Tiene una periodicidad semestral y anual.

El **modelo 0729** tiene como objetivo conocer en unidades físicas y valor (moneda total), la distribución mayorista total y sus destinos, ya sean de bienes de consumo y/o bienes intermedios, así como su existencia final al cierre de cada semestre. Comprende entre otras a las empresas mayoristas centrales y/o universales, empresas mayoristas territoriales de bienes de consumo, empresas que acopian y/o distribuyen productos agrícolas, empresas productoras que realizan la función de circuladoras mayoristas y entidades distribuidoras de las corporaciones CIMEX (siglas de Comercio Interior, Mercado Exterior), CUBALSE (Cuba al Servicio del Extranjero), entre otras.

2.2 Necesidades de información

Con el objetivo de conocer las necesidades de los usuarios finales se realizó un profundo estudio del negocio, mediante encuentros y entrevistas con los especialistas del área de industria y más

específicamente con los encargados de los procesos de distribución. Luego de la realización del análisis se determinó que las principales necesidades de información son las referentes a los totales de distribución por semestre, destino, producto, provincias, los destinos físicos y valor, así como el inventario por semestre, producto y provincia.

2.3 Especificación de requisitos

Con el objetivo de conocer las necesidades del cliente, las características que debe tener el producto y las funciones que debe cumplir el sistema se identifican los requisitos informativos, funcionales y no funcionales.

2.3.1 Requisitos de información

Los requisitos de información son las necesidades básicas de los usuarios y deben estar disponibles en todo momento con el objetivo de brindar una mejora en el proceso de toma de decisiones. A continuación se enuncian los requisitos informativos identificados en la solución:

- Obtener del modelo distribución mayorista de productos seleccionados el total de distribución físico en cuc por provincia, destino, producto y semestre.
- Obtener del modelo distribución mayorista de productos seleccionados el total de distribución físico en cup por provincia, destino, producto y semestre.
- Obtener del modelo distribución mayorista de productos seleccionados el total de distribución físico por provincia, destino, producto y semestre.
- Obtener del modelo distribución mayorista de productos seleccionados el total de distribución valor en cuc por provincia, destino, producto y semestre.
- Obtener del modelo distribución mayorista de productos seleccionados el total de distribución valor en cup por provincia, destino, producto y semestre.
- Obtener del modelo distribución mayorista de productos seleccionados el total de distribución valor por provincia, destino, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el total destino físico en cuc por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el total destino físico en cup por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el total destino físico por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el total destino valor en cuc por provincia, producto y semestre.

- Obtener de los indicadores de la industria manufacturera el total destino valor en cup por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el total destino valor por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el inventario final físico en cuc por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el inventario final físico en cup por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el inventario final físico por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el inventario final valor en cuc por provincia, producto y semestre
- Obtener de los indicadores de la industria manufacturera el inventario final valor en cup por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el inventario final valor por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el real hasta este mes por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el real del año anterior por provincia, producto y semestre.
- Obtener de los indicadores de la industria manufacturera el total destino físico en cuc por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el total destino físico en cup por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el total destino físico por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el total destino valor en cuc por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el total destino valor en cup por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el total destino valor por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el inventario final físico en cuc por producto y semestre.

- Obtener de los indicadores de la industria manufacturera el inventario final físico en cup por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el inventario final físico por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el inventario final valor en cuc por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el inventario final valor en cup por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el inventario final valor por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el real hasta este mes por producto y semestre.
- Obtener de los indicadores de la industria manufacturera el real del año anterior por producto y semestre.

2.3.2 Requisitos funcionales

Los requerimientos funcionales establecen capacidades y funcionalidades que debe cumplir el sistema con el objetivo de satisfacer las necesidades de los usuarios. A continuación se enuncian los requisitos identificados:

- Autenticar usuario.
- Adicionar roles.
- Eliminar roles.
- Adicionar usuario.
- Eliminar usuario.
- Insertar reporte.
- Modificar reporte.
- Eliminar reporte.
- Extraer información.
- Realizar transformación y carga.
- Abrir navegador OLAP.
- Mostrar editor MDX.
- Mostrar padres.
- Ocultar repeticiones.
- Intercambiar ejes.

- Mostrar gráfico.
- Configurar gráfico.
- Configurar impresión.
- Exportar a PDF.
- Exportar a Excel.
- Mostrar propiedades.
- Suprimir filas.
- Detallar miembros.
- Entrar en detalles.
- Mostrar datos de origen.

2.3.3 Requisitos no funcionales

Los requisitos no funcionales especifican criterios que se utilizan para juzgar la operación de un sistema. Una vez realizada la etapa de análisis del negocio se identificaron 26 requisitos no funcionales los cuales quedan plasmados en el artefacto ¨ DATEC-SIGOBDistInd-0113_ERSv1.1¨ y se dividen en las siguientes clasificaciones:

- Requisitos de usabilidad.
- Confiabilidad.
- Requisitos de eficiencia.
- Requisitos de soporte.
- Restricciones de diseño.
- Requisitos para la documentación de usuarios en línea y ayuda del sistema.
- Requisitos de interfaz.
- Requisitos Legales, de Derecho de Autor y otros.

Ejemplo:

Requisitos de usabilidad

RNF 1. 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.

Confiabilidad

RNF 6. 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.

Restricciones del diseño

RNF 15. 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

2.4 Reglas del negocio

Se definen como reglas del negocio el conjunto de normas, políticas, operaciones o restricciones que se establecen con el objetivo de regular los aspectos generales del negocio. En el análisis del negocio se identificaron las siguientes reglas:

- Las cifras deben de poseer un valor numérico.
- En el modelo no puede existir ningún campo vacío.
- Cada código de fila se encuentra asociado a un código de indicador único.
- El total físico de un indicador está dado por la suma del físico en cuc más el físico en cup.
- El total del valor de un indicador está dado por el valor cuc más el valor cup.

2.5 Casos de uso del sistema

Con el objetivo de identificar los casos de uso del sistema se realizó una agrupación de los requisitos funcionales e informativos. De aquí surgen los casos de uso informativos y del sistema, además se identifican los actores del mismo. En el expediente de proyecto definido para la realización de MD se encuentra el artefacto “DATEC-SIGOBDistInd-0114_ECUv1.1” donde quedan plasmados los casos de uso y su descripción.

2.5.1 Actores del sistema

Tabla 2. Actores del sistema.

Actor	Objetivo
Administrador	Administrar los reportes realizados así como los usuarios y otorgar los diferentes roles.
Especialista	Realizar los procesos de inteligencia de negocio que hacen posible la presentación de la información agrupada en dos diferentes casos de uso: Distribución y Balance Semestral. Así como

	autenticarse como usuario con permisos en el sistema.
Administrador ETL	Realizar los procesos de extracción de los datos, transformación y carga.

2.5.2 Especificación de casos de uso del sistema

Durante el análisis del negocio surgieron siete casos de uso funcionales y dos casos de uso informativos, a continuación se enuncian respectivamente:

- Autenticar usuario.
- Administrar Usuario.
- Administrar Reporte.
- Administrar Roles.
- Extraer Datos.
- Realizar transformación y carga.
- Realizar opciones sobre el reporte.
- Presentar el balance semestral de los procesos de producción y distribución de la industria manufacturera.
- Presentar la Distribución de las empresas mayoristas de la industria manufacturera.

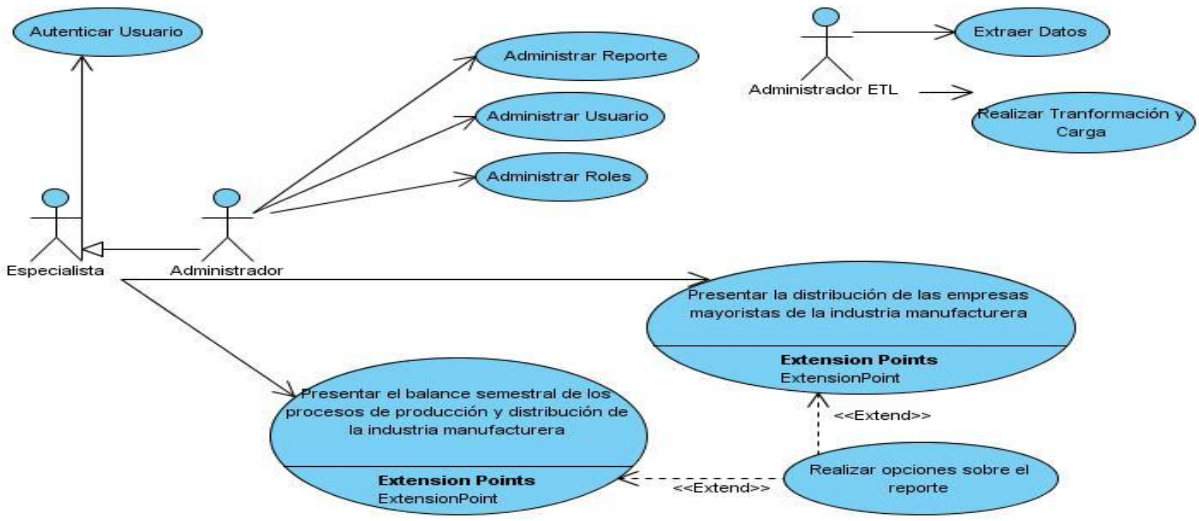


Fig. 1. Diagrama de Casos de Uso del Sistema.

Tabla 3. Descripción del caso de uso Presentar la distribución de las empresas mayoristas de la industria manufacturera.

Objetivo	Presentar indicadores de diferentes áreas.	
Actores	Especialista: (Inicia) Presentar el balance semestral de los procesos de producción y distribución de la industria manufacturera, Presentar la distribución de las empresas mayoristas de la industria manufacturera.	
Resumen	El caso de uso inicia cuando el especialista desea consultar la información referente a los procesos de distribución de las empresas mayoristas de la industria manufacturera. Tiene como objetivo fundamental mostrar los datos solicitados y finaliza cuando el especialista termina con el análisis de la información.	
Complejidad	Media.	
Prioridad	Crítica.	
Precondiciones	El especialista tiene que estar autenticado. El AD tiene que estar poblado.	
Pos condiciones	Los reportes correspondientes fueron consultados por el especialista.	
Flujo de eventos		
Flujo básico < Presentar la distribución de las empresas mayoristas de la industria manufacturera >		
	Actor	**Sistema**
1.	El usuario selecciona A.A G SIGOB.	
2.		Muestra las áreas de análisis correspondientes.
3.	Selecciona AA_Distribución_Industria_Manufacturera.	
4.		Muestra LT_Distribución_Empresas_Mayoristas.
5.	Selecciona el Reporte.	
6.		Visualiza el Reporte. Se brindan las opciones del usuario. Ir al CU Realizar opciones de reporte.

7.		Termina el caso de uso.
Flujos alternos		
2a Introduce los datos incorrectamente		
	Actor	**Sistema**
		Muestra mensaje ¨Los datos son incorrectos¨. Se vuelve al paso uno del flujo básico de eventos.

Opciones de Reporte de Presentar la distribución de las empresas mayoristas de la industria manufacturera

Perspectivas de análisis	**Posibles resultados**	
	Medidas	**Periodicidad**
Variables de entrada relacionadas con el CU Presentar la distribución de las empresas mayoristas del área industria manufacturera: • Provincia • Destino • Producto • Semestre	Variables de salida disponibles en el hecho Distribución: • total_distrib_fisico_cuc • total_distrib_fisico_cup • total_distrib_fisico • total_distrib_valor_cuc • total_distrib_valor_cup • total_distrib_valor	Rango de tiempo en que se solicitan las variables de salida: • Semestral

Relaciones	**CU Incluidos**	No aplica.
	CU Extendidos	Realizar opciones de reporte: paso ocho del Flujo Básico. Realizar opciones de reporte en el CU Presentar Indicadores.
Requisitos no funcionales	Sección: “3.2 Requisitos no funcionales” del documento: “0114_Especificación de requisitos de software”.	
Asuntos pendientes	Posibles mejoras al caso de uso.	

2.6 Arquitectura del sistema

La arquitectura del software es uno de los aspectos más importantes que deben tenerse en cuenta en la construcción de un AD. A través de la arquitectura del software se puede representar la estructura que tendrá el sistema. En el caso de los almacenes de datos la arquitectura es la forma en que se representan todas las estructuras de datos, comunicaciones y procesos.

Para tener una visión general del sistema se explica a continuación la arquitectura de la solución propuesta, detallando cada uno de los subsistemas que la conforman. Se identificaron tres subsistemas en los cuales el sistema está estructurado: Subsistema de integración, Subsistema de almacenamiento y Subsistema de visualización.

Fig. 2. Arquitectura del sistema.

- **Subsistema de integración:** se agrupan los procesos encargados de llevar a cabo tareas relacionadas con la extracción, integración y limpieza de los datos, además de la carga y actualización del MD. Para el desarrollo de estos procesos se hace uso las herramientas DataCleaner y Pentaho Data Integration.
- **Subsistema de almacenamiento:** es el encargado de contener toda la información correspondiente al MD. Este estará compuesto por dimensiones y tablas de hechos que a su vez contendrán los datos que describirán un hecho así como las medidas. Para el almacenamiento de la información se hace uso del gestor de base de datos PostgreSQL y como aplicación gráfica para la administración de la base de datos PgAdmin.
- **Subsistema de visualización:** es la vista de presentación y tiene como finalidad mostrar los datos almacenados de forma útil a través de las herramientas Pentaho BI Server y Web Apache Tomcat; es la capa con la cual interactúa el usuario final. Se comunica directamente con el servidor OLAP a través de consultas realizadas mediante las herramientas Mondrian OLAP Server y Schema Workbench.

2.7 Diseño de la solución

El diseño de la solución se basa en las etapas por las que transita el desarrollo del MD, el diseño del subsistema de almacenamiento, de integración y por último el diseño del subsistema de visualización.

2.7.1 Diseño del subsistema de almacenamiento

- **Tablas de dimensiones**

Dimensión destino: la dimensión destino describe los valores por los cuales se clasifica la información haciendo referencia a los destinos de los productos.

Dimensión indicador general: la dimensión indicador general describe los valores por los cuales se clasifica la información haciendo referencia a los indicadores de los productos seleccionados.

Dimensión temporal semestre: la dimensión temporal semestre describe los valores correspondientes a los semestres y al año al que corresponde.

Dimensión temporal provincia: la dimensión temporal provincia contiene la información correspondiente a las provincias de todo el país.

- **Tablas de hechos**

Hecho distribución: la tabla hecho distribución contiene la información referente al proceso de distribución tanto en valor como en físico, en todas las monedas: cuc, cup y en moneda total, por provincia, destino, indicador y semestre.

Hecho balance semestral: la tabla hecho balance semestral contiene la información referente a los totales de los destinos físico y valor en cuc, cup y moneda total. Almacena también el inventario final en físico, valor y en todas las monedas así como el real de este mes y el real del año anterior por provincia, indicador y semestre.

Matriz BUS

La matriz BUS o matriz dimensional se confecciona con el objetivo de conocer cómo quedarán representadas las relaciones entre los hechos y las dimensiones identificadas. A continuación se muestra la matriz bus realizada para el MD:

Tabla 4. Matriz BUS.

Hecho/Dimensión	Provincia	Indicador	Destino	Semestre
Distribución	X	X	X	X
Balance semestral	X	X		X

Modelo de datos

El modelo de datos tiene como objetivo fundamental brindar la estructura de los datos y está conformado por hechos, dimensiones y medidas. La topología utilizada en el MD es constelación de hechos debido a que existen dimensiones compartidas entre las tablas de hechos. La figura tres representa el modelo de datos del MD Distribución del área industria manufacturera donde quedan representados los hechos identificados con sus medidas y relaciones correspondientes.

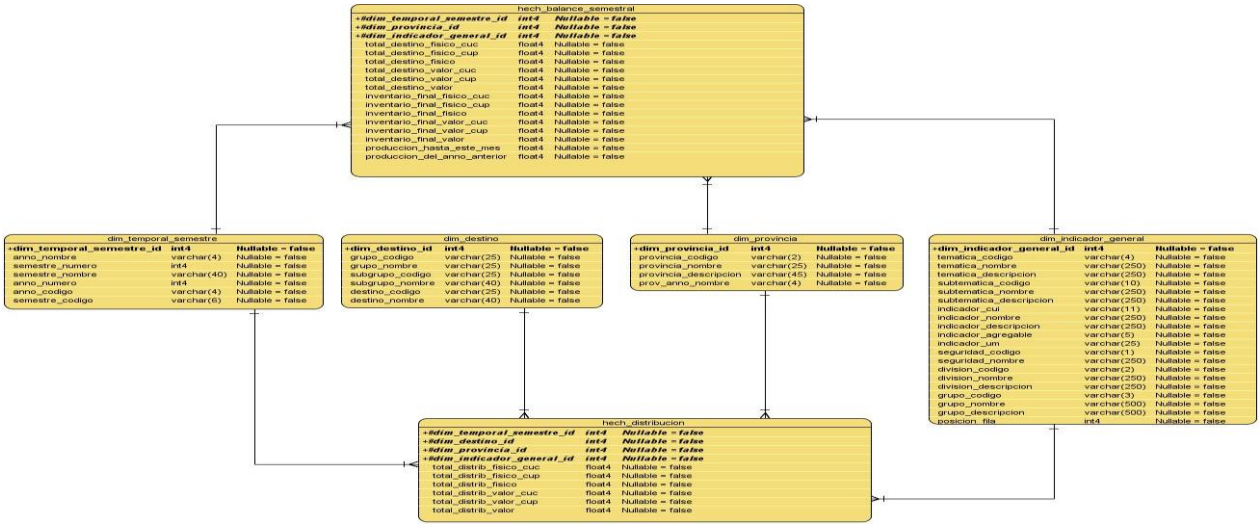


Fig 3. Modelo de datos.

2.7.2 Diseño del subsistema de integración

Perfilado de datos

Por perfilado de datos se entiende el análisis de los datos de los sistemas fuentes para entender su contenido, estructura, calidad y dependencias. A partir de este proceso se establecen reglas para corregir los defectos de los datos y así garantizar la disponibilidad de los mismos. Para el análisis de los datos correspondiente a los procesos de distribución del área industria manufacturera se utilizó la herramienta DataCleaner con la que se pudo concluir la cantidad de elementos total por cada una de

las columnas de las tablas fuentes así como los elementos distintos, nulos, duplicados, mínimos y máximos.

Diseño del subsistema de integración

Los procesos de ETL suponen una gran parte del tiempo empleado en el desarrollo de un AD. Para la correcta integración de los datos se realiza el diseño de los procesos de integración. Estos constituyen un conjunto de pasos lógicos que al ser ejecutados permiten que sean cargados los datos en las tablas diseñadas en el modelo dimensional. En la figura cuatro se observa el diseño general seguido para la implementación de las transformaciones con respecto a la fuente incremental:

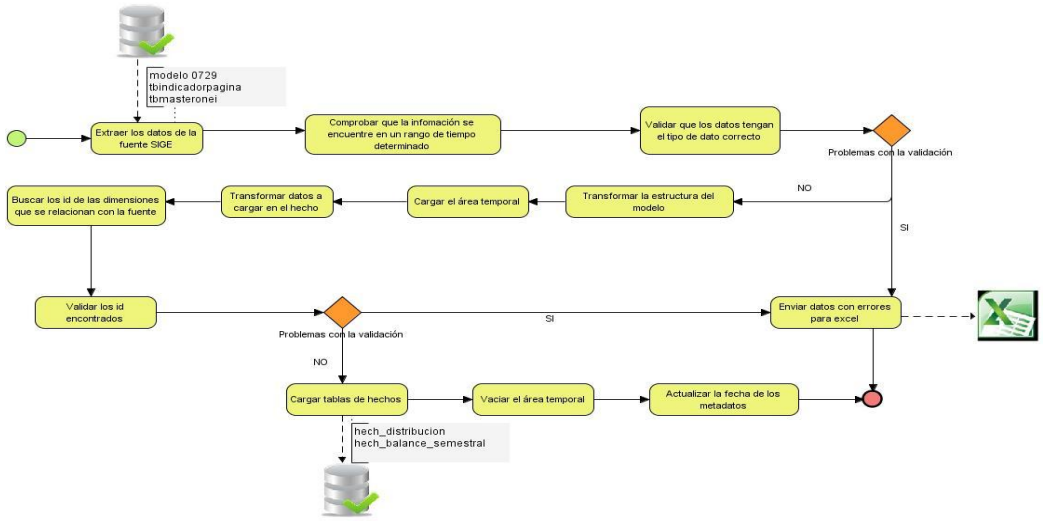


Fig. 4. Diseño del subsistema de integración.

Registro de sistemas fuentes

El expediente de proyecto contiene una serie de artefactos que ayudan a la comprensión de todo lo relacionado con el MD Distribución del área industria manufacturera para SIGOB, uno de los artefactos es el Registro de Sistemas Fuentes. Este describe detalladamente cada una de las fuentes de datos correspondientes al MD, enfocado principalmente en la ubicación física de los datos.

2.7.3 Diseño del subsistema de visualización

Arquitectura de información

El artefacto arquitectura de información tiene como objetivo fundamental facilitar el análisis, control y monitoreo de la información del proceso de distribución del sector de industria en la ONEI para el apoyo al proceso de toma decisiones. Se identificó un Área de Análisis (A.A) que contiene dos libros de trabajo (LT) y diecisiete tablas de salida agrupadas en dichos LT.

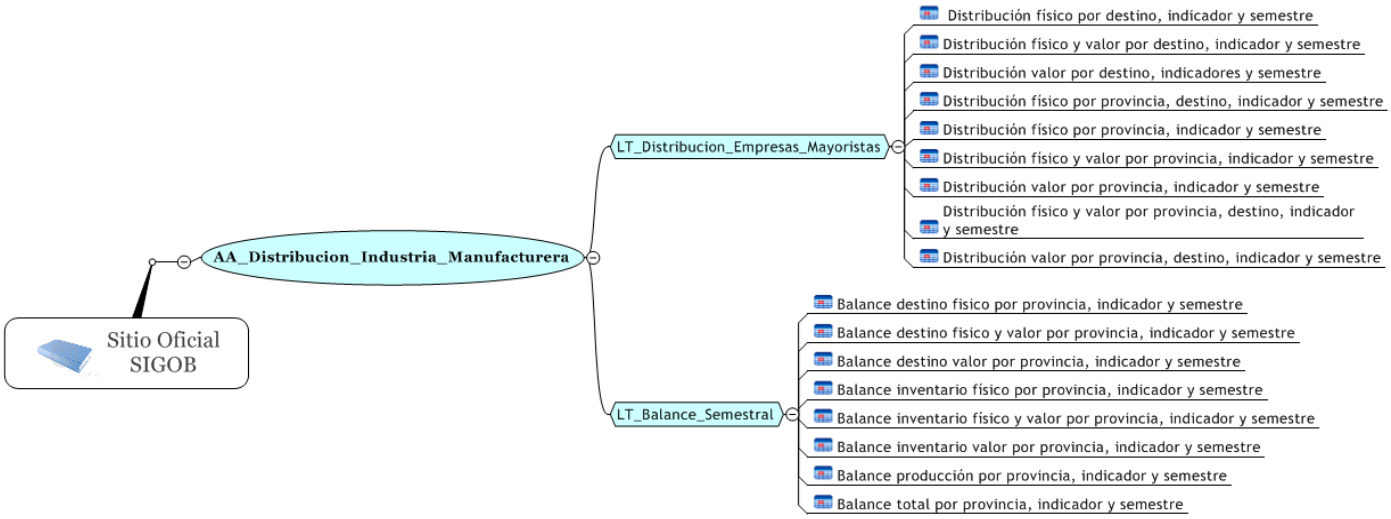


Fig. 5. Mapa de Navegación.

Diseño de los cubos OLAP

Uno de los factores claves en el procesamiento analítico en línea son los cubos OLAP, estos proveen rápido acceso a los datos almacenados independientemente de la cantidad de datos en el cubo, al mismo tiempo son un subconjunto de datos del AD. Para el diseño de los cubos OLAP se utilizó la herramienta Schema Workbench. Se modelaron un total de dos cubos en correspondencia con cada tabla de hecho y cuatro dimensiones con sus respectivos niveles jerárquicos; además el esquema cuenta con 20 medidas enfocadas todas hacia las necesidades del cliente.

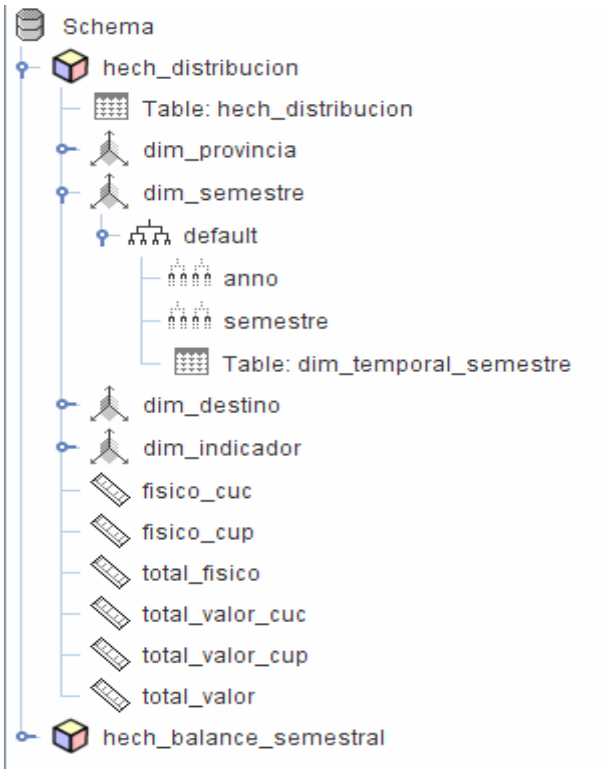


Fig. 6. Cubos OLAP.

Diseño de los reportes candidatos

El artefacto Reportes Candidatos describe los reportes que fueron identificados en el desarrollo del MD Distribución del área industria manufacturera para de esta forma poder comprender los elementos que los componen. La figura siete muestra la descripción de uno de ellos.

Área de análisis (AA)	Distribución Industria Manufacturera
Libro de Trabajo (LT)	LT1– Balance_Semestral
Reporte (Tabla de Salida – TS)	TS2– Balance destino valor por provincia, indicador y semestre
Descripción	El reporte muestra el balance semestral de los indicadores seleccionados por producto y semestre.
Elementos del reporte	➢ Provincia ➢ Indicador ➢ Semestre
Frecuencia de emisión	Semestral

Fig. 7. Tabla de reporte candidato.

2.8 Política de respaldo y recuperación

Los aspectos fundamentales en los que se puede medir la política de respaldo y recuperación que se utiliza en el MD Distribución del área industria manufacturera para el AD SIGOB son:

- **Periodicidad de las salvas:** se realizan salvas de la base de datos con una periodicidad semestral.
- **Tablas involucradas:** la tabla de hecho balance semestral y distribución mayorista.
- **Salvas existentes:** no existen.

2.9 Esquema de seguridad

Los niveles de acceso al sistema garantizan la seguridad del MD Distribución del área industria manufacturera para el AD SIGOB, todo radica en los permisos y roles que tienen asignados los usuarios para acceder e interactuar con la base de datos y la aplicación. Con motivo de brindar mayor seguridad a la base de datos se definió el rol de Administrador que tiene control total sobre la misma y además se definieron los siguientes roles para la seguridad de la aplicación:

Tabla 5. Roles y permisos.

Roles	Permisos
Administrador de base de datos	Tiene acceso total a la base de datos, siendo el responsable de conceder los permisos para el acceso a esta y además se encarga del respaldo y recuperación de la base de datos.
Administrado de ETL	Se encarga de gestionar todos los procesos de ETL.
Administrador	Tiene acceso total a todas las Áreas de Análisis General (AAG). Gestiona el Sistema de Información de Gobierno.
Analista	Tiene acceso de solo lectura al Áreas de Análisis (AA) Distribución del área industria manufacturera para SIGOB.

Tabla 6. Nivel de acceso a los elementos de aplicación.

Elementos de aplicación	Roles con acceso
AA General	Administrador
Carpeta raíz: AA Distribución del área industria manufacturera para SIGOB.	Administrador y Analista

2.10 Conclusiones

Una vez realizado el análisis y diseño del MD Distribución del área industria manufacturera para el AD SIGOB se obtuvieron las conclusiones siguientes:

- Se definieron los requisitos informativos, funcionales y no funcionales que permitirán el desarrollo de la solución, cumpliendo con las necesidades del cliente.
- Se definieron cinco reglas del negocio como base para las reglas de transformación aplicadas en los procesos de ETL.
- Se describió la arquitectura base del MD detallando los elementos que conforman el sistema.
- Se realizó el diseño de los subsistemas de almacenamiento, integración y visualización de datos que representan la guía base para la implementación del MD.

Capítulo 3. Implementación del mercado de datos Distribución del área industria manufacturera

En el presente capítulo se aborda lo referente a la implementación de los procesos de almacenamiento, integración y visualización de los datos, con el propósito de brindar una mejor comprensión de las estrategias y procedimientos utilizados, mostrándose las potencialidades de las herramientas seleccionadas en el capítulo uno. Una vez concluida la implementación del modelo de datos físico del MD, se procede a la carga de los datos mediante los procesos de ETL hacia su destino y finalmente se realiza la automatización de todo el proceso de visualización de la información.

3.1 Implementación del subsistema de almacenamiento

La correcta implementación del subsistema de almacenamiento es de vital importancia, pues es así como quedará estructurado el MD. Durante este proceso se definen los esquemas del mercado y los estándares de codificación del mismo que garantizarán la correcta integración con el AD SIGOB.

3.1.1 Estructura de los datos

El desarrollo exitoso de un modelo físico es un aspecto importante dentro de la construcción de un AD. El punto de partida para llevarlo a cabo es el modelo lógico, el cual no es más que el modelo de datos dimensional representado en el capítulo anterior. Para la implementación del modelo de datos físico se utilizó el gestor de base de datos PostgreSQL. A continuación se describe la implementación del modelo de datos físico en el MD quedando planteada la estructura de la base de datos.

Esquemas

Los esquemas constituyen la manera de organizar de forma lógica las tablas en la base de datos. En el MD distribución se definieron cinco esquemas que se describen a continuación:

- **Esquema dimensiones**: contiene todas las dimensiones comunes del AD.
- **Esquema correlacionadores:** contiene las tablas que almacenan los correlacionadores de los modelos con el objetivo de obtener el código de indicador.
- **Esquema mart_industria_distribución:** contiene las tablas de hechos y las dimensiones propias del MD.
- **Esquema sgt_etl:** esquema que contiene las tablas temporales que permitirán la carga de los hechos.

- **Esquema metadatos:** contiene las tablas de metadatos que almacenan la fecha en que fue insertada la información por última vez y los metadatos para observar el comportamiento que tienen las transformaciones y trabajos al ser ejecutados.

Tablas

El MD Distribución cuenta con un total de 13 tablas distribuidas en los esquemas anteriormente mencionados. De estas 13 tablas dos son hechos y cuatro dimensiones, de las cuales dos de ellas son propias del mercado.

Tabla 7. Esquemas y tablas.

Esquemas	Tablas
dimensiones	dim_provincia
	dim_temporal_semestre
mart_industria_distribución	hech_distribución
	hech_balance_semestral
	dim_destino
	dim_indicador_general
correlacionadores	correlacionador_0006
	correlacionador_0729
sgt_etl	mart_industria_distribución_modelo_0729
	mart_industria_distribución_balance_semestre_0729
metadatos	meta_distind_trans
	meta_distind_job
	md_traza_carga

3.1.2 Estándares de codificación

Con el objetivo de mantener la estandarización con el AD SIGOB se establecieron un conjunto de reglas y normas. De esta forma se determinó nombrar los componentes creados de la siguiente manera:

Tabla 8. Estándares de codificación.

Tipo de objeto	Función	Nomenclatura	Descripción
BD	dwh	[nombre]_dwh	Almacén de datos
Esquema	Dimensiones compartidas	Dimensiones	Esquemas donde se organizan las dimensiones compartidas por varios MD (tablas y secuencias).
	Esquemas de datos	mart_[temática]	Esquemas donde se almacenan las tablas de hechos y tablas necesarias para gestionar los datos asociados a cada área temática.
Tablas	Dimensiones	dim_[nombre]	Tablas dimensionales utilizadas.
	Hechos	hech_[nombre]	Tablas de hechos que definen las principales medidas requeridas para calcular indicadores y otras medidas derivadas.
Secuencias	Secuencias dimensiones	seqDim_[tabla][nombre]	Secuencias dimensionales.

3.2 Implementación del subsistema de integración

Con el objetivo de cargar satisfactoriamente el MD se llevó a cabo la implementación del subsistema de integración. Durante este proceso se le realizaron un conjunto de operaciones a los datos para garantizar la calidad e integridad de los mismos.

3.2.1 Subsistema de extracción

Los procesos de extracción se encargan de obtener los datos provenientes de las múltiples fuentes y cargarlos hacia el área temporal donde recibirán el tratamiento que permitirá que sean cargados posteriormente a las tablas de hecho. Se identificaron dos tipos de fuentes, la fuente proveniente de los archivos con extensión dbf que son consideradas históricas debido a que solo se almacena la información perteneciente a los años anteriores al 2011 y la fuente incremental que es la información que se encuentra almacenada en la base de datos Sistema Integrado de Gestión Estadística (SIGE) que contiene la información referente a los años posteriores al 2011 la cual es actualizada semestralmente.

3.2.2 Limpieza y transformación de datos

Después de realizada la extracción de los datos se procede a limpiar los mismos. Debido al gran cúmulo de información los datos pueden contener errores, estar duplicados o ser incoherentes, para lo que es necesario realizarles un adecuado tratamiento con el objetivo de obtener la información precisa. Los procesos de transformación y limpieza de la información son muy importantes y de la correcta realización de estos depende la calidad que tendrán los datos que se mostrarán al usuario final. En este proceso se aplican las reglas del negocio anteriormente expuestas.

3.2.3 Transformaciones y trabajos

Con el objetivo de poblar el MD se realizaron un total de 20 transformaciones. Cuatro de estas transformaciones tienen como objetivo la carga de los hechos. Para la carga la fuente histórica se realizó un total de cinco transformaciones y para la fuente incremental 13. Debido a la extensión de la solución, a continuación solo se explicarán las transformaciones realizadas para cargar las tablas de hecho con la fuente incremental.

- **Transformaciones para la carga de las dimensiones**

carga_dim_indicador: transformación que carga la dimensión indicador general. Primeramente se extraen los campos necesarios de las tablas tbindindicador, tbtematica, tbclasificaciondeseguridad y tbindicadorpagina, todas ubicadas en la base de datos SIGE. Luego de ordenar los campos se hacen coincidir con la división y el grupo provenientes del Excel. Finalmente se carga la dimensión indicador general la cual debe encontrarse actualizada siempre que se carguen los hechos semestralmente.

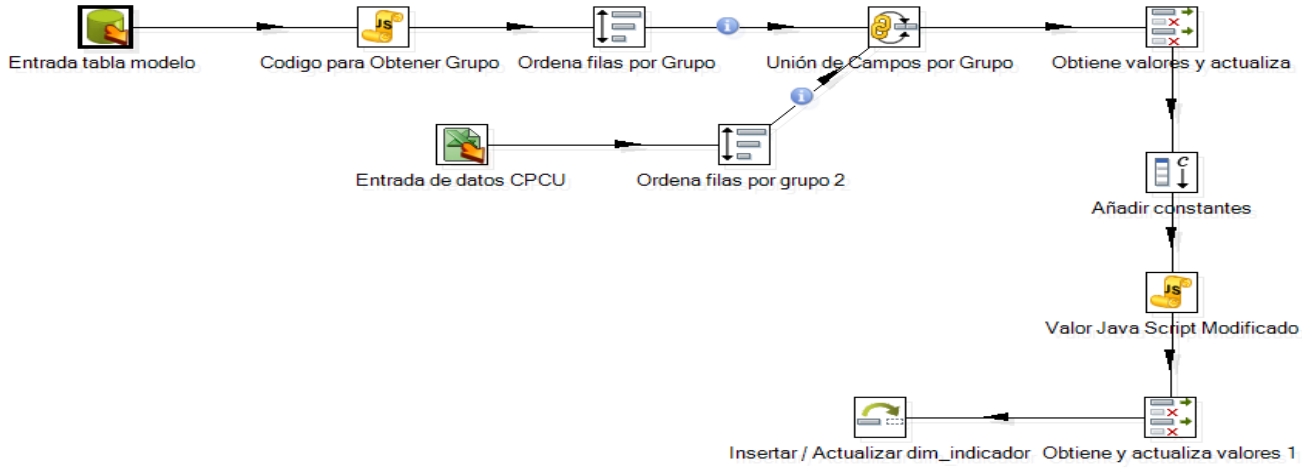


Fig. 8. Carga de la dimensión indicador general.

carga_dim_destino: transformación que carga la dimensión destino. Primeramente se extraen los datos almacenados en el Excel para que sean cargados en la dimensión destino.

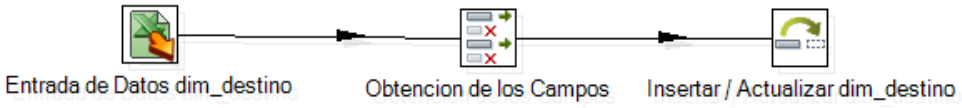


Fig. 9. Carga de la dimensión destino.

- **Transformaciones para el llenado de los hechos**

cargar_sgt_modelo_0729_sige: transformación que carga el área temporal con los datos provenientes de la fuente de datos SIGE. Primeramente se seleccionan los datos a utilizar del modelo 0729, siempre que se encuentre en un rango de fecha entre la última vez que se cargó el mercado y la fecha más actual que se encuentre en la fuente. Con los campos del modelo seleccionados se procede a obtener, todavía de la base de datos fuente, la distribución política administrativa (DPA) para a partir de esta obtener la provincia. Posteriormente se realizan los cambios necesarios a los datos y se agrupan las filas de manera que mediante el paso Java Script se puedan separar los 13 destinos provenientes del modelo en 52 columnas, de forma tal que de la columna uno a la 13 se obtenga los destinos en físico cup, de la 14 a la 26 el valor en cup, de la 27 a la 39 el físico en cuc y por último de

la 40 a la 52 el valor en cuc. El área temporal tiene que estar vacía para garantizar que los datos que se carguen en las tablas de hecho sean referentes a un rango de fecha determinado.

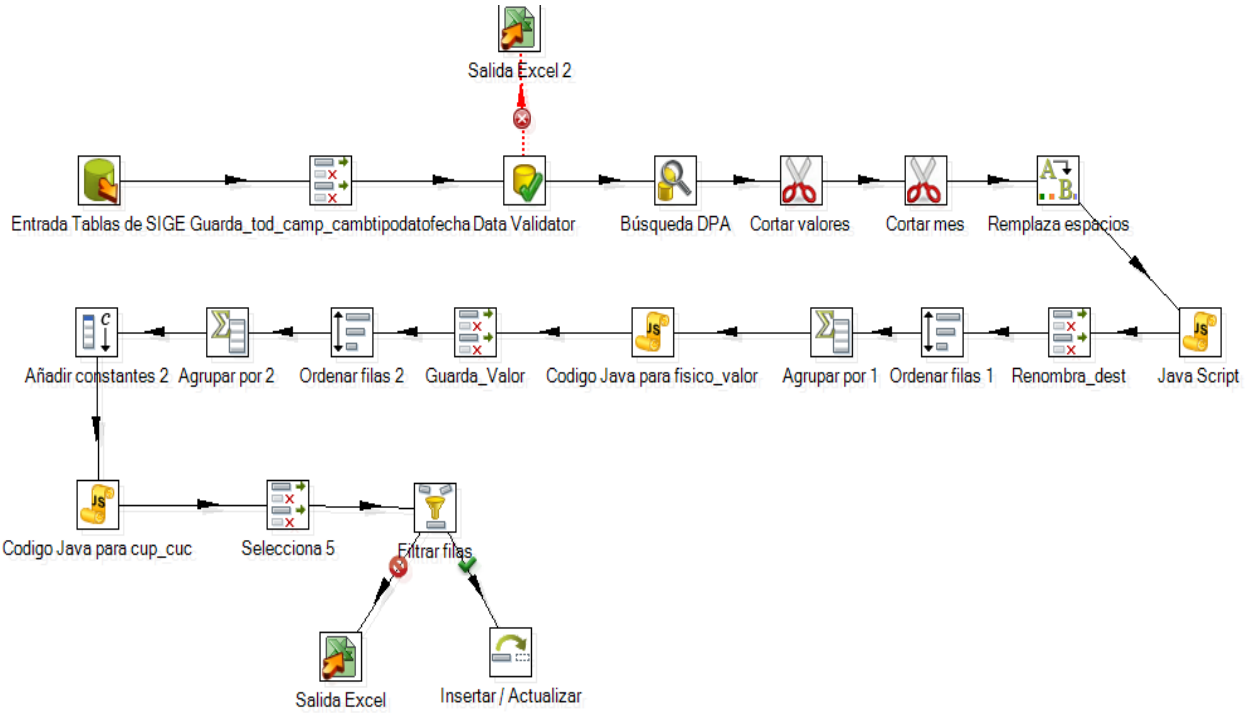


Fig. 10. Carga del área temporal.

carga_hech_balance_semestral_sige: transformación que carga el hecho balance semestral con los datos provenientes de la fuente de datos SIGE. Primeramente se seleccionan los campos que se necesitan del área temporal. Posteriormente se calculan los totales del destino y el inventario. Se realizan una serie de transformaciones a los datos y se obtiene mediante el código de fila el código del indicador de la tabla indicador página proveniente de la fuente. Finalmente, luego de ordenar y agrupar los campos se almacenan en la segunda tabla temporal donde se encuentren los campos del destino total en todas sus variantes y el inventario.

En la segunda entrada de tabla se obtienen los datos provenientes del modelo 0006. Se realiza una búsqueda del DPA para obtener a través de este la provincia. Posteriormente se procede a obtener el código del indicador a partir del código de fila y seguidamente se realiza una búsqueda en la tabla temporal que se cargó anteriormente, donde se hace coincidir el código del semestre y el código del indicador para de esta manera correlacionar los campos provenientes del modelo 0006 y el 0729. Por último se realizan todas las búsquedas en la base de datos para encontrar los identificadores de las dimensiones que se relacionan con este hecho.

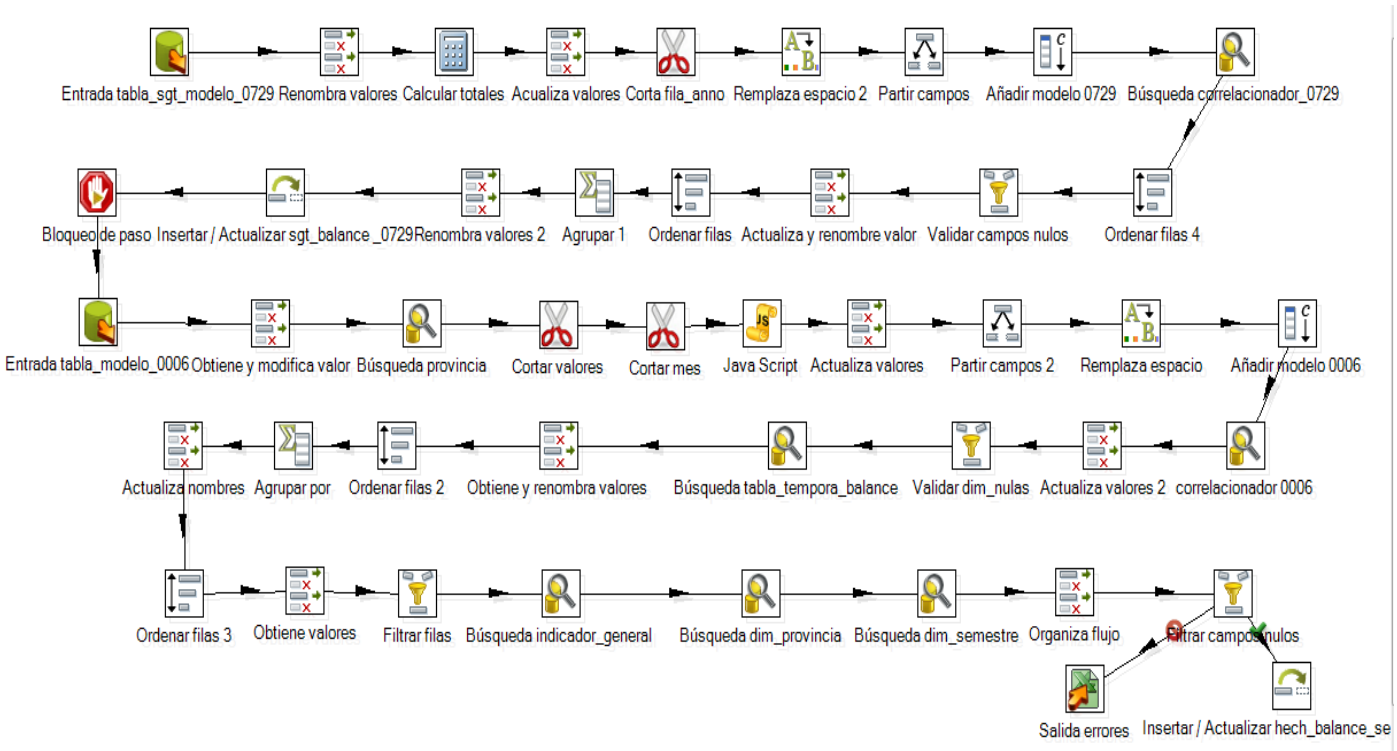


Fig. 11. Carga del hecho balance semestral.

carga_hech_distribución_sige: transformación que carga el hecho distribución con los datos proveniente de la fuente SIGE. Primeramente se seleccionan los campos a utilizar del área temporal ya cargada, que serán las 52 columnas pertenecientes a los 13 destinos en todas sus variantes. Posteriormente se calculan los totales de los destinos para normalizar los campos de manera que las 52 columnas queden agrupadas en seis, físico en todas sus variantes y valor en todas sus variantes. Una vez realizado este paso se procede a efectuar las búsquedas en la base de datos para encontrar los identificadores de las dimensiones que se relacionan con este hecho.

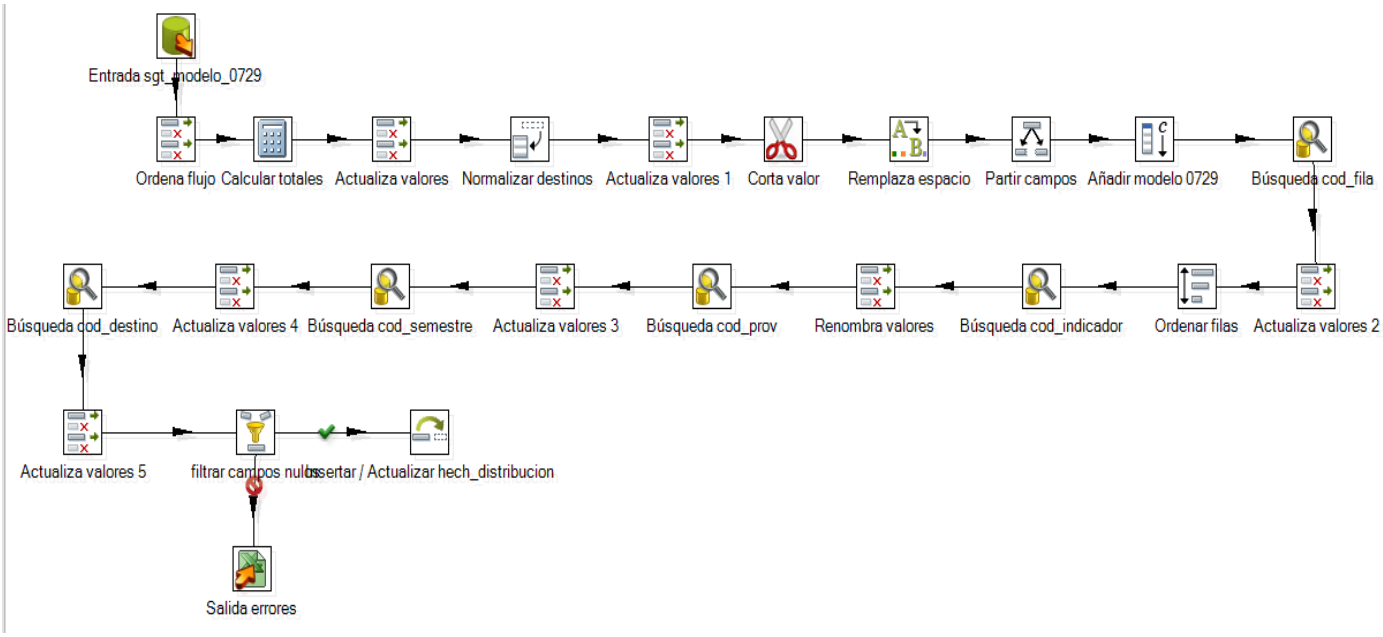


Fig. 12. Carga de la tabla de hecho distribución.

- **Trabajos**

Una vez que se han realizado todas las transformaciones necesarias, es preciso organizar el proceso de carga al MD, para esto se realizan los trabajos, los cuales están pensados para controlar la ejecución e interacción de las transformaciones. De forma general se realizaron cuatro trabajos que cargan los hechos referentes a la fuente histórica e incremental y uno principal que realiza la ejecución de todos los demás. A continuación se explican de manera general la implementación de los trabajos:

trab_principal_cargar_hechos_sige: trabajo que ejecuta las transformaciones referentes a la carga de los hechos de la fuente incremental. También ejecuta las transformaciones que se encargan de actualizar la fecha de las tablas de metadatos y borrar el área temporal.

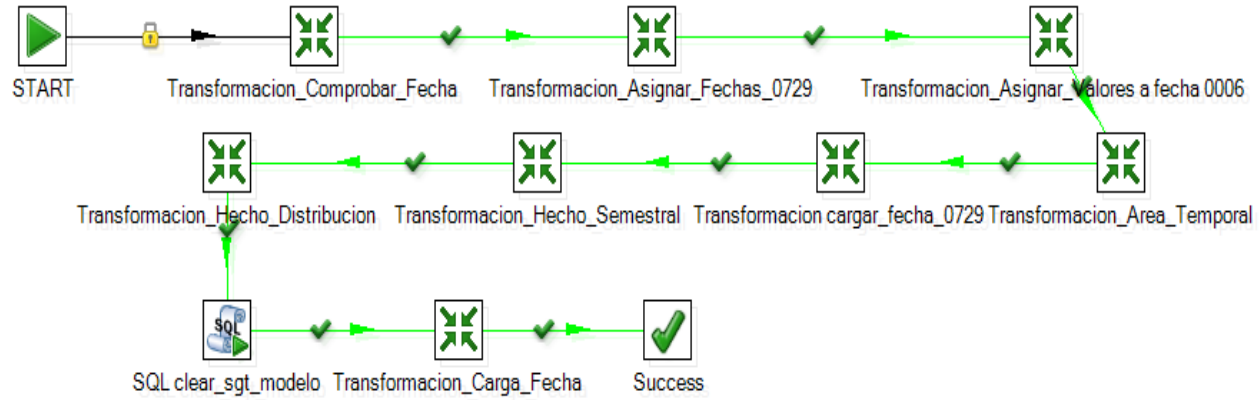


Fig. 13. Trabajo que ejecuta las transformaciones para cargar hechos SIGE

MD_Distribución_Industria: trabajo que ejecuta los trabajos para cargar los hechos y las dimensiones propias del MD con las fuentes tanto incremental como histórica.

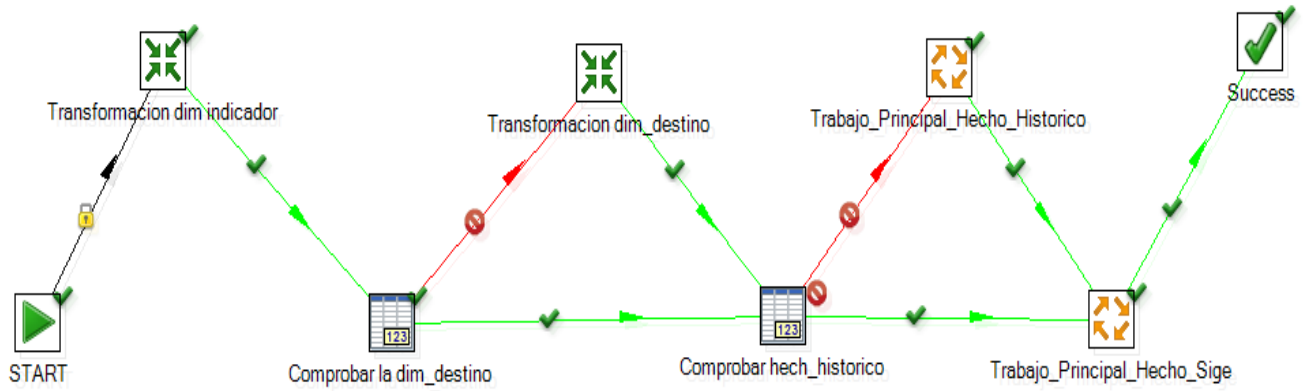


Fig. 14. Trabajo principal del MD.

3.2.4 Carga de datos

La carga consiste en el proceso de almacenar los datos en el MD final una vez realizada la extracción y transformación de los mismos. La carga referente a la fuente histórica mencionada anteriormente se efectúa solamente la primera vez que se carga el MD, mientras que las cargas incrementales se realizan semestralmente. Estas cargas incrementales garantizan que la información siempre se encuentre actualizada.

3.2.5 Gestión del cambio en las dimensiones

Se le conoce como Dimensiones Lentamente Cambiantes (*Slowly Changing Dimensions*, SCD) a las dimensiones cuyos datos varían con el paso del tiempo, ya sea de forma ocasional o constante, o que implique a un solo registro o la tabla completa. Cuando ocurren estos cambios, se puede optar por seguir alguna de estas dos opciones: registrar el historial de cambios o reemplazar los valores que sean necesarios. Existen métodos para tratar este problema como son los del tipo 0, 1, 2, 3, 4 y 6 enunciados a continuación: (19)

Tipo 0: este es un enfoque pasivo, es decir no se hace nada al respecto. Los valores permanecen como estaba la dimensión cuando los registros fueron creados.

Tipo 1: en este enfoque se sobrescriben los datos viejos con el dato actualizado sin mantener el historial de donde perteneció. Este enfoque es sencillo, pero tiene la desventaja de que no contiene historial de los datos.

Tipo 2: en este enfoque se inserta un nuevo registro cada vez que existe un cambio en la dimensión. Se agrega un campo de versión u opcionalmente se agregan dos columnas para capturar la fecha de inicio y final de ese valor.

Tipo 3: este método da seguimiento al cambio agregando nuevas columnas. Este enfoque solo puede mantener un cambio histórico.

Tipo 4: este método mantiene una tabla histórica para todos los cambios y una tabla con el valor actual de la dimensión.

Tipo 6 Híbrido: este método es una combinación de los tipos 1,2 y 3 (1 + 2 + 3 = 6). El enfoque es usar una dimensión tipo 1 (escribiendo el dato actual), pero agregar un par adicional de columnas con las fechas de validez (Tipo 2).

En el desarrollo del MD se identificó la dimensión indicador general SCD de tipo 1, debido a que sus datos pueden actualizarse con el paso del tiempo. Otras de las características de la dimensión indicador general es no es necesario mantener un registro de la información histórica.

3.2.6 Gestión de los metadatos del proceso de integración

Uno de los componentes más importantes de la arquitectura de un AD son los metadatos. Se definen comúnmente como "datos acerca de datos". Según Ralph Kimball los metadatos se pueden dividir en tres categorías que se describen a continuación:

- Los metadatos técnicos que proporcionan una traza de las actividades y objetos del AD.
- Los metadatos del negocio que proporcionan un contexto para interpretar los datos, descripciones de negocio de un elemento de datos, informaciones sobre cuando fue cargado el

dato y transformado, todo con el objetivo de que el usuario pueda entender los datos y usarlos para la toma de decisiones.

- Los metadatos de procesos se refieren a los datos que se actualizan siempre que se ejecute un proceso.

En la realización del MD se identificaron los metadatos de procesos, debido a que se almacenan los datos pertenecientes a las fechas correspondientes a la última vez que se insertó la información en la fuente de datos. También se recogen los datos que permiten conocer el comportamiento que tienen las transformaciones y trabajos al ejecutarse. Se crearon tres tablas de metadatos que se describen a continuación:

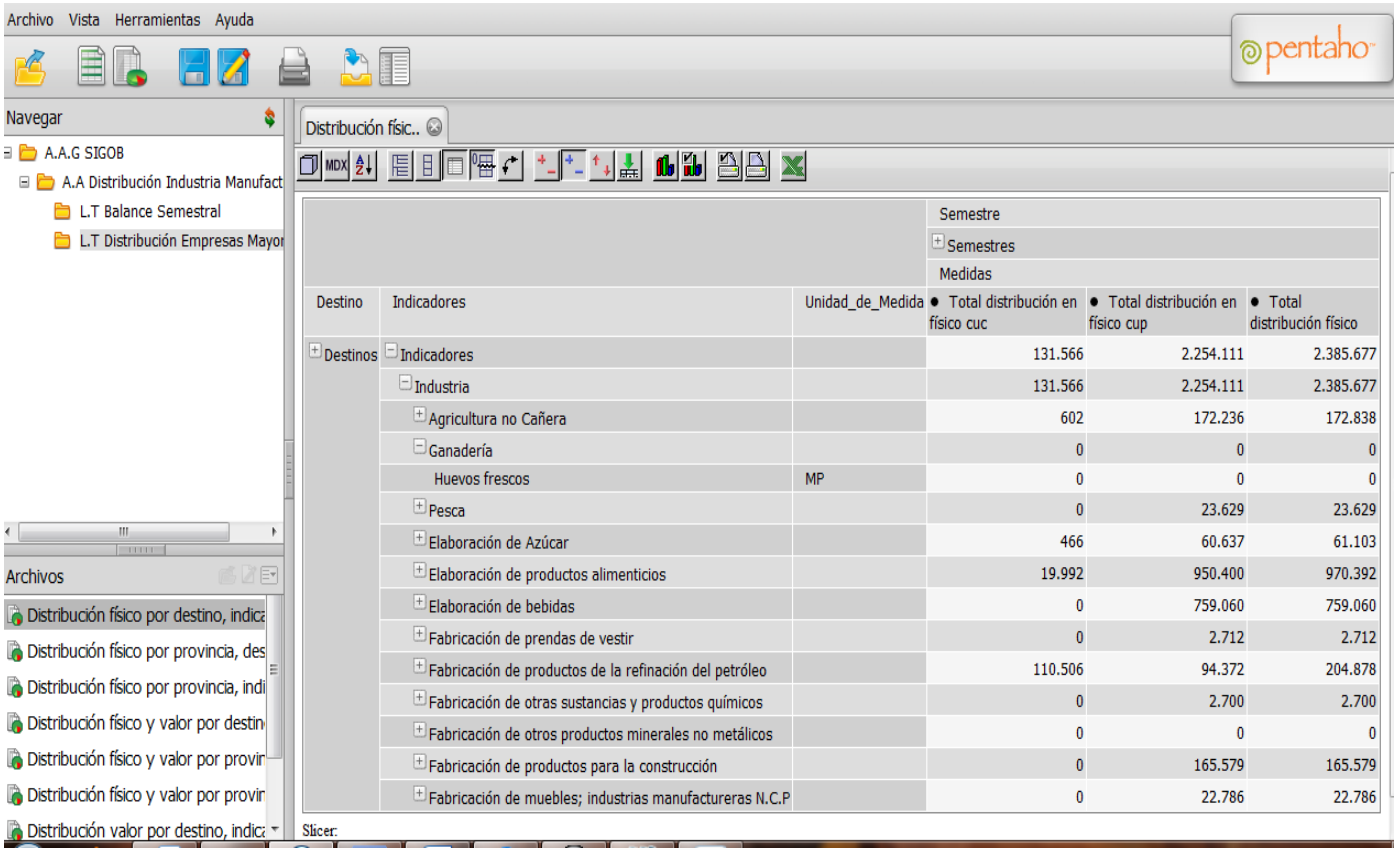
- **md_traza_carga:** almacena la fecha correspondiente a la última vez que se cargó el MD y todos sus datos asociados como la fuente de datos, el modelo y el nombre del MD al que corresponde la información.
- **meta_distind_trans:** guarda la información que permite conocer la cantidad de datos que fueron leídos, escritos, actualizados, además se puede seguir el estado de una transformación en tiempo de ejecución.
- **meta_distind_job:** tabla de metadatos que almacena el estado con respecto a los trabajos al ejecutarse, también brinda la información referente a la cantidad de líneas leídas, escritas, actualizadas y los posibles errores que puedan existir.

3.3 Implementación del subsistema de visualización de datos

Con el objetivo de mostrar la información requerida por el cliente se lleva a cabo la implementación del subsistema de visualización. Durante este proceso se realizan un conjunto de tareas para que el usuario final pueda analizar de manera eficiente la información contenida en el sistema.

3.3.1 Implementación de la capa de visualización

La solución propuesta cuenta con 17 vistas de análisis referentes a las 17 tablas de salida y distribuidas en sus correspondientes L.T que tienen como objetivo satisfacer las necesidades de información de los usuarios. El Pentaho BI-Server permite mostrar el resultado del análisis realizado, a través de esta aplicación es posible ver los reportes y analizar la información mediante tablas de datos o gráficas de diversos tipos (barras, pastel, líneas); permitiendo además, desplegar el cubo de información y modificar las vistas de análisis o crear unas totalmente nuevas. Por otra parte, esta interfaz brinda la posibilidad de cambiar el tipo de gráfica, imprimir el reporte o salvarlo en un archivo de formato PDF o XLS. La figura 15 muestra una vista de análisis correspondiente al libro de trabajo Distribución de Empresas Mayoristas:

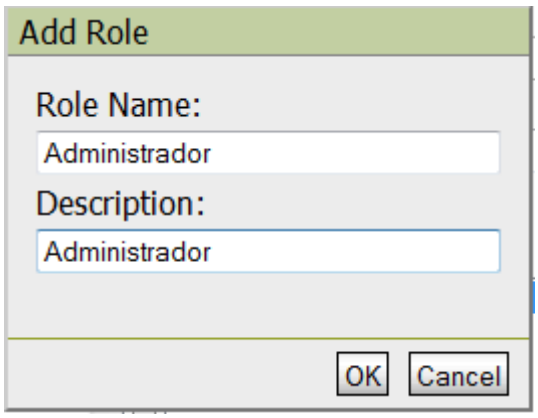


			Semestre		
			Semestres		
			Medidas		
Destino	Indicadores	Unidad_de_Medida	• Total distribución en físico cuc	• Total distribución en físico cup	• Total distribución físico
Destinos	Indicadores		131.566	2.254.111	2.385.677
	Industria		131.566	2.254.111	2.385.677
	Agricultura no Cañera		602	172.236	172.838
	Ganadería		0	0	0
	Huevos frescos	MP	0	0	0
	Pesca		0	23.629	23.629
	Elaboración de Azúcar		466	60.637	61.103
	Elaboración de productos alimenticios		19.992	950.400	970.392
	Elaboración de bebidas		0	759.060	759.060
	Fabricación de prendas de vestir		0	2.712	2.712
	Fabricación de productos de la refinación del petróleo		110.506	94.372	204.878
	Fabricación de otras sustancias y productos químicos		0	2.700	2.700
	Fabricación de otros productos minerales no metálicos		0	0	0
	Fabricación de productos para la construcción		0	165.579	165.579
	Fabricación de muebles; industrias manufactureras N.C.P		0	22.786	22.786

Fig. 15. Vista de análisis.

3.3.2 Configurar la seguridad de los usuarios y roles

Durante la implementación del subsistema de visualización del MD Distribución del área industria manufacturera se crearon dos usuarios los cuales tienen diferentes permisos para acceder a la información, proporcionando una mayor seguridad al sistema. El rol de administrador tiene todos los permisos de la aplicación y se corresponde con el usuario administrador del sistema. El rol de analista tiene permiso de solo lectura y se corresponde a su vez con el usuario analista del sistema.



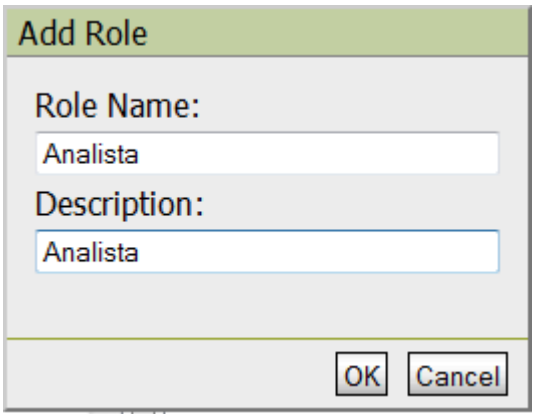


Fig. 16. Roles.

3.4 Conclusiones

Durante el desarrollo del capítulo se describió el modelo de datos físico, la implementación de los subsistemas de integración y visualización, así como los pasos para la implantación del sistema. Una vez realizada la implementación del MD se obtuvieron los resultados siguientes:

- Se implementó el subsistema de almacenamiento de los datos detallándose la estructura física del MD obteniéndose como resultado la distribución de las tablas correspondientes al mercado con respecto a los esquemas que conforman el mismo.
- Se implementó el subsistema de integración de los datos logrando cargar el MD satisfactoriamente y cumpliendo a su vez con las reglas de transformación.
- Se implementó el subsistema de visualización de los datos obteniéndose como resultado las vistas correspondientes a los LT identificados, permitiendo estas la visualización de la información.

Capítulo 4. Validación del mercado de datos Distribución del área industria manufacturera

Una vez desarrollado el análisis, diseño e implementación del MD es necesario validar su funcionamiento y comprobar el éxito del mismo. Para cumplir este objetivo se realizan los casos de prueba por cada caso de uso de información y reglas de transformación, las listas de chequeo para los procesos de ETL y el perfilado de los datos para la información almacenada en el MD final.

4.1 Pruebas

Las pruebas se centran principalmente en la evaluación o la valoración de la calidad del producto y representan un elemento crítico para la garantía del mismo. Es una actividad en la cual un sistema o uno de sus componentes se ejecutan en circunstancias previamente especificadas, los resultados se observan, se registran y se realiza una evaluación de algún aspecto. El objetivo de la etapa de pruebas es garantizar la calidad del producto desarrollado. Además, esta etapa implica:

- Verificar el funcionamiento de los componentes.
- Verificar la integración adecuada de los componentes.
- Verificar que todos los requisitos se han implementado correctamente.

Existen un conjunto de posibles pruebas que se le pueden realizar a un sistema informático para validar su uso. Para lograr obtener un producto con calidad que cumpla con los requerimientos establecidos y la satisfacción del cliente, se realizaron las pruebas de software de acuerdo al Modelo V definido por CALISOFT e implementado por el centro DATEC. Este modelo V considera las actividades de prueba como un proceso que se ejecuta en paralelo con las actividades de análisis y diseño, en lugar de establecer una fase independiente al final del proyecto. La figura 17 representa el ciclo de vida del software propuesto en el Modelo V, donde a la izquierda se muestran las diferentes etapas de desarrollo, mientras que a la derecha se observan las pruebas correspondientes a cada una de ellas.

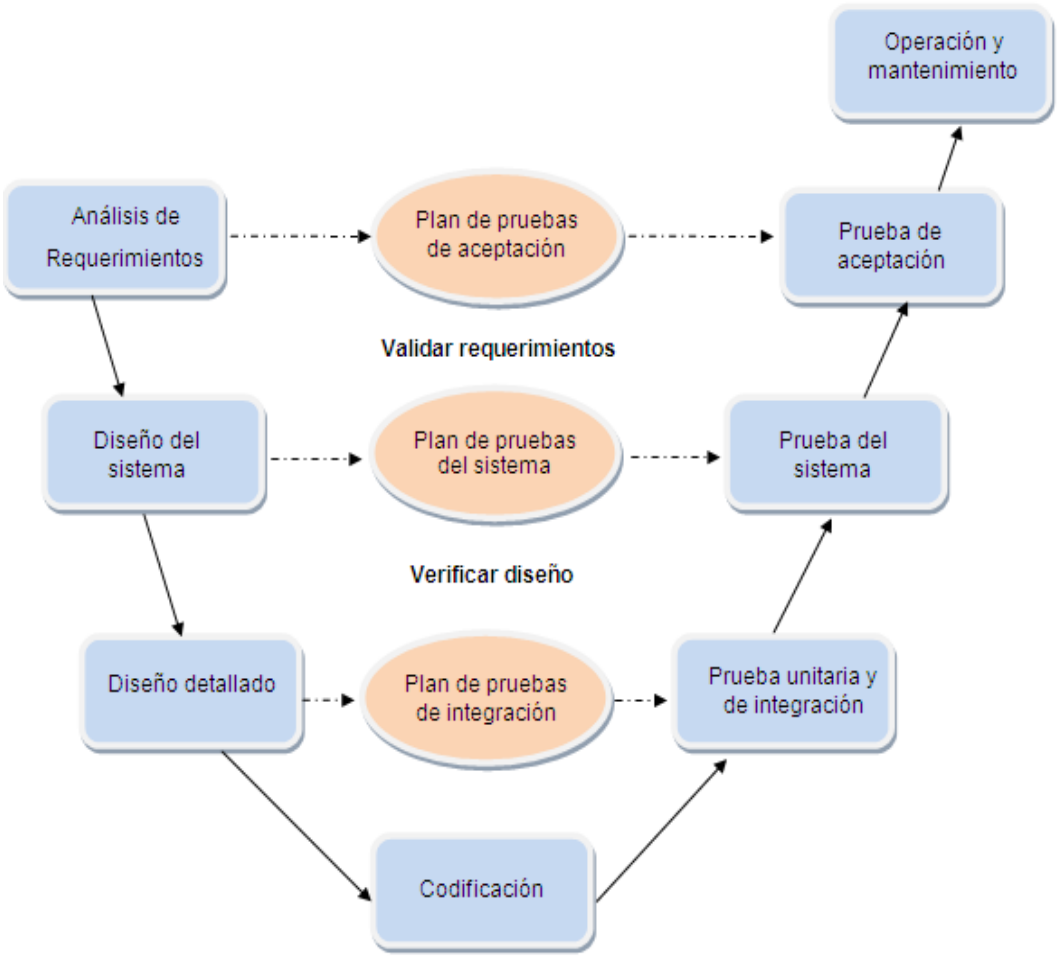


Fig. 17. Modelo V.

A continuación se exponen algunas de las pruebas que pueden ser utilizadas para la validación de un producto de software: (20)

- **Prueba unitaria:** es el proceso de probar los componentes individuales de la solución. El propósito es identificar diferencias entre la especificación de los artefactos y el comportamiento real de cada módulo.
- **Prueba de integración:** consiste en construir el sistema a partir de los distintos componentes y probarlo con todos integrados. Estas pruebas deben realizarse progresivamente.
- **Prueba del sistema:** se refiere al comportamiento del sistema integrado. La prueba del sistema se aplica generalmente para probar los requerimientos no funcionales de la solución.
- **Pruebas de aceptación:** se realizan para probar que el sistema cumpla con los requerimientos especificados por el cliente.

4.1.1 Casos de prueba

El propósito de un caso de prueba es especificar una forma de probar el sistema, incluyendo las entradas para la validación, los resultados esperados y las condiciones bajo las que ha de probarse. Para validar los requerimientos del sistema de la investigación se le realizan casos de prueba a cada caso de uso informativo, con el objetivo de comprobar la disponibilidad de los perfiles de análisis y los indicadores a medir, así como también verificar el cumplimiento de los requisitos de información a través de los reportes candidatos. Se diseñaron dos casos de prueba basados en casos de uso informativos los cuales se encuentran detallados en los artefactos generados por cada caso de uso (casos de prueba basados en casos de uso). Por otra parte se diseñaron los casos de pruebas basados en las reglas de transformación definidas en el análisis de la solución.

4.2 Listas de chequeo

Las listas de chequeo son un conjunto de preguntas que se elaboran en forma de cuestionario con el objetivo de verificar el grado de cumplimiento de ciertas reglas establecidas con un fin determinado. Las preguntas en forma de cuestionario sirven como una guía que obliga a quienes las contestan a reflexionar sobre el nivel de cumplimiento de determinados requisitos. Las listas de chequeo enumeran una serie de *ítems* que deberían verificarse uno a uno para asegurarse de lograr el producto final con un nivel de calidad previamente aceptado. Las listas de chequeo definidas para la evaluación de la calidad del MD están divididas en tres secciones:

- **Estructura del documento:** abarca todos los aspectos definidos por el expediente de proyecto o el formato establecido por el proyecto**.**
- **Indicadores definidos en el desarrollo:** abarca todos los indicadores a evaluar durante la etapa de ETL.
- **Semántica del documento**: contempla todos los indicadores a evaluar respecto a la ortografía, redacción, entre otros aspectos.

Elementos que forman parte de la estructura de la lista de chequeo:

- **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 la etapa.
- **Evaluación (Eval):** es la forma de evaluar el indicador en cuestión. El mismo se evalúa de uno en caso de que exista alguna dificultad sobre el indicador y de cero en caso de que el indicador revisado no presente problemas.
- **No Procede (N.P):** se usa para especificar que no es necesario evaluar el indicador.

- **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. Pueden o no existir señalamientos o sugerencias.

Evaluación a través de la lista de chequeo:

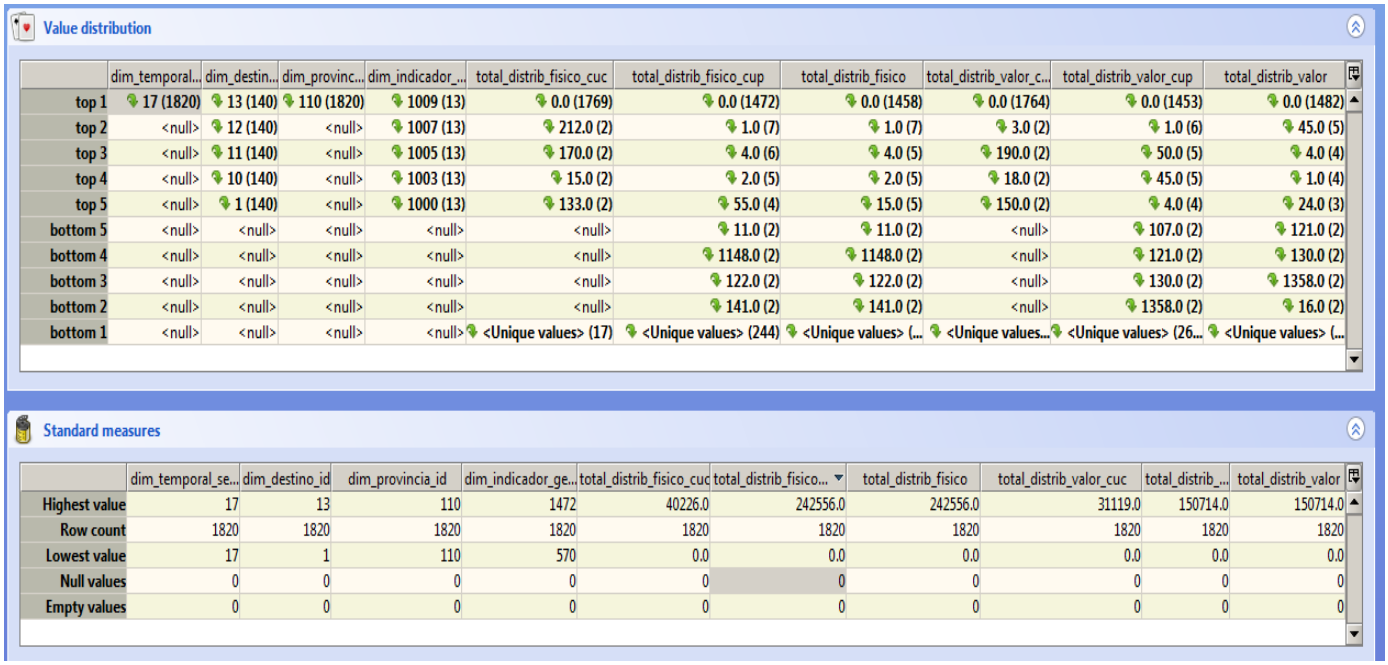
Tabla 9. Lista de chequeo.

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?				
crítico	2. ¿El alcance del proyecto describe correctamente los datos de las dimensiones y hechos del mercado de datos?				
crítico	3. ¿El objetivo expresa correctamente el propósito del documento?				
	4. ¿Se hace un uso adecuado del control del documento?				
	5. ¿En la sección de acrónimos se definen todos los acrónimos utilizados en el documento?				
	6. ¿En el entregable, la definición de las variables se hace correctamente?				

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

4.3 Calidad de los datos

Con el objetivo de conocer la calidad de los datos almacenados en el MD se realizó el perfilado de los mismos a través de la herramienta DataCleaner. Una vez analizado el estado de estos se concluyó que no existen valores nulos en el destino, también se definieron la cantidad de valores únicos, duplicados, distintos y los mínimos y máximos de cada medida de las tablas de hechos. La figura 18 muestra el perfilado de los datos realizado al hecho distribución.



	dim_temporal...	dim_destin...	dim_provinc...	dim_indicador_...	total_distrib_fisico_cuc	total_distrib_fisico_cup	total_distrib_fisico	total_distrib_valor_c...	total_distrib_valor_cup	total_distrib_valor
top 1	17 (1820)	13 (140)	110 (1820)	1009 (13)	0.0 (1769)	0.0 (1472)	0.0 (1458)	0.0 (1764)	0.0 (1453)	0.0 (1482)
top 2	<null>	12 (140)	<null>	1007 (13)	212.0 (2)	1.0 (7)	1.0 (7)	3.0 (2)	1.0 (6)	45.0 (5)
top 3	<null>	11 (140)	<null>	1005 (13)	170.0 (2)	4.0 (6)	4.0 (5)	190.0 (2)	50.0 (5)	4.0 (4)
top 4	<null>	10 (140)	<null>	1003 (13)	15.0 (2)	2.0 (5)	2.0 (5)	18.0 (2)	45.0 (5)	1.0 (4)
top 5	<null>	1 (140)	<null>	1000 (13)	133.0 (2)	55.0 (4)	15.0 (5)	150.0 (2)	4.0 (4)	24.0 (3)
bottom 5	<null>	<null>	<null>	<null>	<null>	11.0 (2)	11.0 (2)	<null>	107.0 (2)	121.0 (2)
bottom 4	<null>	<null>	<null>	<null>	<null>	1148.0 (2)	1148.0 (2)	<null>	121.0 (2)	130.0 (2)
bottom 3	<null>	<null>	<null>	<null>	<null>	122.0 (2)	122.0 (2)	<null>	130.0 (2)	1358.0 (2)
bottom 2	<null>	<null>	<null>	<null>	<null>	141.0 (2)	141.0 (2)	<null>	1358.0 (2)	16.0 (2)
bottom 1	<null>	<null>	<null>	<null>	<Unique values> (17)	<Unique values> (244)	<Unique values> (...	<Unique values...	<Unique values> (26...	<Unique values> (...

	dim_temporal_se...	dim_destino_id	dim_provincia_id	dim_indicador_ge...	total_distrib_fisico_cuc	total_distrib_fisico...	total_distrib_fisico	total_distrib_valor_cuc	total_distrib_...	total_distrib_valor
Highest value	17	13	110	1472	40226.0	242556.0	242556.0	31119.0	150714.0	150714.0
Row count	1820	1820	1820	1820	1820	1820	1820	1820	1820	1820
Lowest value	17	1	110	570	0.0	0.0	0.0	0.0	0.0	0.0
Null values	0	0	0	0	0	0	0	0	0	0
Empty values	0	0	0	0	0	0	0	0	0	0

Fig. 18. Perfilado de datos.

Una vez concluidas las pruebas de calidad de los datos el cliente probó el funcionamiento y las funcionalidades del MD para verificar que este cumpliera con los requisitos establecidos. El usuario final determinó que la solución satisface las necesidades de información anteriormente definidas. De esta forma el cliente confirmó la aceptación del MD Distribución del área industria manufacturera para SIGOB.

4.4 Evaluación del resultado de las pruebas

Aplicación de los casos de prueba

Los casos de prueba se realizaron con el objetivo de especificar una manera de probar el sistema. La aplicación de los casos de prueba basados en casos de uso llevadas a cabo en el centro DATEC, dieron como resultado un total de cuatro no conformidades, las cuales fueron resueltas satisfactoriamente. Por otra parte, la aplicación de los casos de prueba para las reglas de transformación no arrojó no conformidades.

Aplicación de las listas de chequeo

Evaluación

Se aborta la revisión si:

- Existen al menos dos indicadores críticos evaluados de mal en la sección **Indicadores** que posee la lista de chequeo.
- Más del 50 % de los indicadores a evaluar están evaluados de mal.
- Se mantienen las no conformidades de una revisión a otra.

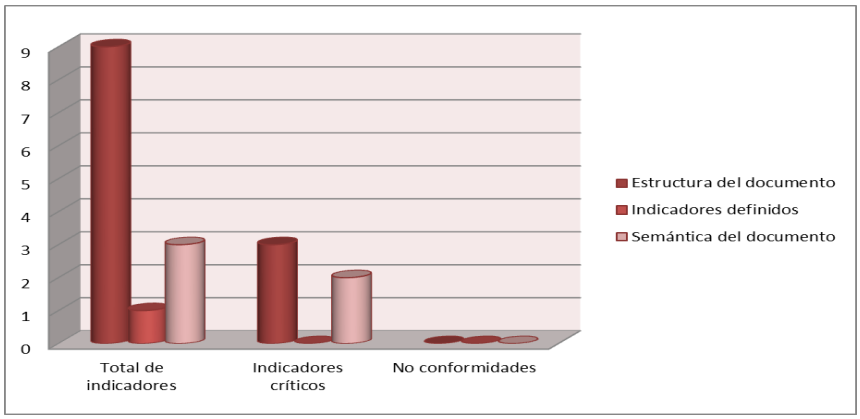
Se evalúa de regular la calidad del diseño revisado si no cumple los criterios para ser abortado y además:

- Incumple con los indicadores críticos a evaluar de las secciones **Estructura del documento** y **Semántica del documento** que posee la lista de chequeo.
- Existe al menos un indicador crítico evaluado de mal.
- Existen al menos cinco indicadores no críticos evaluados de mal de la sección **Indicadores evaluados por la etapa** que posee la lista de chequeo.

El diseño es evaluado de bien si no cumple ninguno de los criterios anteriores y además:

- No existe ningún indicador crítico evaluado de mal.
- Si la cantidad de indicadores no críticos evaluados de mal de la sección **Indicadores** que posee la lista de chequeo no es mayor que cuatro.

En la figura 19 se observa el comportamiento de los indicadores definidos para la lista de chequeo elaborada para el artefacto Diccionario de Datos perteneciente a los procesos de ETL. De forma general se identificaron 13 indicadores, de ellos cinco críticos y luego de aplicada la lista de chequeo no se generaron no conformidades. Durante la realización de las listas de chequeo para los procesos de ETL no se identificaron no conformidades.



- **Fig. 19. Comportamiento de los indicadores por secciones.**

4.5 Conclusiones

En este capítulo se expuso la validación del MD, obteniéndose los siguientes resultados:

- Se diseñaron y aplicaron los casos de prueba diseñados, verificando que el MD cumple con la calidad requerida.
- Se diseñaron y aplicaron las listas de chequeo a los procesos de ETL, validando y garantizando que este proceso cumple con las especificaciones requeridas.
- Se realizó el perfilado de los datos al MD Distribución del área industria manufacturera comprobando el estado de los datos en el MD.

Conclusiones

La realización del presente trabajo de diploma abordó el estudio del estado actual de los almacenes de datos y la importancia que tiene la utilización de los mismos para el proceso de toma de decisiones. Con la elaboración del trabajo se cumplieron los objetivos trazados:

- Se fundamentó la metodología, herramientas y tecnologías utilizadas durante el desarrollo del MD. Se utilizó la Metodología para el desarrollo de soluciones de almacenes de datos e inteligencia de negocio propuesta por DATEC orientando el proceso de desarrollo del MD. Se definieron las herramientas a utilizar para el diseño y la implementación del MD.
- Se realizó el análisis y diseño del MD Distribución del área industria manufacturera obteniéndose primeramente los requisitos informativos, funcionales y no funcionales, cumpliendo estos con las necesidades de los usuarios finales. Se diseñó el modelo de datos físico y lógico logrando definir la estructura del MD, quedando así planteado el diseño del subsistema de almacenamiento. Se diseñaron los subsistemas de integración y visualización de datos sirviendo como base para la fase de implementación.
- Se implementó el MD Distribución del área industria manufacturera creándose primeramente los esquemas y tablas en el gestor utilizado lo que permitió separar las dimensiones comunes y las tablas de hecho. Se implementó el subsistema de integración logrando cargar satisfactoriamente el MD y se realizó el subsistema de visualización permitiendo la consulta de la información a través de las vistas de análisis.
- Se validó el MD Distribución del área industria manufacturera realizando casos de prueba y listas de chequeo arrojando como resultado la calidad de los datos en el MD y garantizando que la solución cumple con las necesidades de los usuarios.

Recomendaciones

- Desplegar el Mercado de datos Distribución del área industria manufacturera para el Sistema de Información de Gobierno en la Oficina Nacional de Estadísticas e Información.
- Añadir como funcionalidad a la aplicación la posibilidad de analizar la información a través de mapas geo referenciales.

Referencias Bibliográficas

1. **Gil Soto, Esperanza.** Invenia. [En línea] Septiembre de 2001. http://www.invenia.es .

2. **Marrero Antunez, Ivette.** StrateBI. [En línea] Noviembre de 2008. http://www.stratebi.com .

3. [En línea] http://www.sinnexus.com/business_intelligence/datawarehouse.aspx .

4. Sinnexus. [En línea] 2007. http://www.sinnexus.com .

5. [En línea] Revista Ciencias Básicas UJAT, 1 junio 2007. http://www.publicaciones.ujat.mx/publicaciones/revista_dacb/Acervo/v6n1OL/v6n1a5-ol/v6n1a5-ol.html#bib02.

6. ediciona. [En línea] 14 de abril de 2008. http://www.ediciona.com .

7. etl-tools. [En línea] 2006. http://etl-tools.info .

8. Lerrot. [En línea]. http://www.lerrot.com .

9. **Rivadera, Gustavo.** UCASAL Universidad Catolica de Salta. [En línea] 2010. http://www.ucasal.net/templates/unid-academicas/ingenieria/apps/5-p56-rivadera-formateado.pdf .

10. *Propuesta de metodologia para el desarrollo de los almacenes de datos en DATEC,* **Hernandez,Ing. Yanisbel González. 2011**

11. Free Donwload Manager. [En línea] 5 de Marzo de 2007. http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_(Iglesia_Anglicana)_%5BMac_OS_X_cuenta_14717_p/ .

12. PostgreSQL. [En línea] 1996. http://www.postgresql.org/docs/8.4/static/intro-whatis.html, .

13. pgAdmin PostgreSQL Tool. [En línea] 2010. http://www.pgadmin.org/. .

14. DataCleaner. [En línea] 2011. http://datacleaner.eobjects.org/ .

15. Gravitar. [En línea] 2011. http://www.gravitar.biz

16. Mondrian. [En línea] http://mondrian.pentaho.com/documentation/workbench.php .

17. Pentaho. [En línea] 2005. http://www.pentaho.com/explore/products/?hp=y...

18. Apache Tomcat Fundation. [En línea] 2008. http://tomcat.apache.org/...

19. **Díaz, Josep Curto.** *Introducción al business intelligence.* Barcelona: UOC 2010

20. **Sommerville, Ian.** *Ingeniería del Software.* s.l: Prentice Hall, 2005. ISBN: 8478290745.

Bibliografía

1. Apache Tomcat Fundation. [Online] 2008. http://tomcat.apache.org/...
2. DataCleaner. [Online] 2011. http://datacleaner.eobjects.org/.
3. **Díaz, Josep Curto.** *Introducción al business intelligence.* Barcelona: UOC 2010
4. ediciona. [Online] abril 14, 2008. http://www.ediciona.com.
5. etl-tools. [Online] 2006. http://etl-tools.info.
6. Free Donwload Manager. [Online] Marzo 5, 2007. http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_(Iglesia_Anglicana)_%5BMac_OS_X_cuenta_14717_p/ .
7. **Gil Soto, Esperanza.** Invenia. [Online] Septiembre 2001. http://www.invenia.es .
8. Gravitar. [Online] 2011. http://www.gravitar.biz .
9. **Kimball, Ralph.** *The Data Warehouse ETL Toolkit*. Indianapolis: Wiley Publishing, Inc,2004. eISBN:0-764-57923-1
10. Lerrot. [Online] www.lerrot.com .
11. **Marrero Antunez, Ivette.** StrateBI. [Online] Noviembre 2008. http://www.stratebi.com .
12. Mondrian. [Online] http://mondrian.pentaho.com/documentation/workbench.php .
13. Pentaho. [Online] 2005. http://www.pentaho.com/explore/products/?hp=y...
14. pgAdmin PostgreSQL Tool. [Online] 2010. http://www.pgadmin.org/ .
15. PostgreSQL. [Online] 1996. http://www.postgresql.org/docs/8.4/static/intro-whatis.html, .
16. **Rivadera, Gustavo.** UCASAL Universidad Catolica de Salta. [Online] 2010. http://www.ucasal.net/templates/unid-academicas/ingenieria/apps/5-p56-rivadera-formateado.pdf .
17. **Sanchez, Leopoldo Zenaido Zepeda.** Metodología Conceptual para el Desarrollo de los Almacenes de Datos. Valencia : s.n.,2008
18. **Sanz, Miguel Rodríguez.** Universidad Carlos III de Madrid. [Online] Julio 22, 2010. http://www.uc3m.es/portal/page/portal/inicio .
19. Sinnexus. [Online] 2007. http://www.sinnexus.com .
20. **Sommerville, Ian.** *Ingeniería del Software. s.l: Prentice Hall,* 2005. ISBN: 8478290745.

21. [En línea] http://www.sinnexus.com/business_intelligence/datawarehouse.aspx.

22. [En línea] Revista Ciencias Básicas UJAT, 1 junio 2007. http://www.publicaciones.ujat.mx/publicaciones/revista_dacb/Acervo/v6n1OL/v6n1a5-ol/v6n1a5-ol.html#bib02.

GLOSARIO

Almacén de Datos: es una estructura que se define en función de temas específicos, donde la información histórica debe estar integrada y robusta ante los cambios que puedan afectar a la organización. Su objetivo principal, es servir de ayuda a la toma de decisiones empresariales.

Base de datos relacional: es una base de datos que cumple con el modelo relacional, el cual es el modelo más utilizado en la actualidad para implementar bases de datos ya planificadas. Permiten establecer relaciones entre los datos (que están guardados en tablas), y a través de ellas relacionar los datos de ambas tablas, de ahí proviene su nombre: "Modelo Relacional".

BI: inteligencia de negocio. Conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa.

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

CUS: casos de uso del sistema. Proceso dentro del negocio que se estudia, por lo que se corresponde con una secuencia de acciones con un orden lógico, y que producen un resultado observable para ciertos actores del negocio.

DATEC: Centro de Tecnologías de Gestión de Datos.

ETL: Extracción, Transformación y Carga. Proceso que organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un Almacén de Datos, reformatearlos, limpiarlos y cargarlos en otra base de datos, Almacén o Mercado de Datos.

HTTP: Hypertext Transfer Protocol (en español protocolo de transferencia de hipertexto) es el protocolo usado en cada transacción de la World Wide Web.

Mercado de Datos: es una base de datos departamental que se especializa en almacenar datos de un área específica, brindando una estructura óptima para analizar los procesos que tienen lugar dentro del departamento. Son AD orientados a temas específicos y contienen datos de solo una línea del negocio.

OLAP: es el acrónimo en inglés de procesamiento analítico en línea. Es una solución utilizada en el campo de inteligencia de negocio, cuyo objetivo es agilizar la consulta de grandes cantidades de datos.

ONEI: Oficina Nacional de Estadísticas e Información.

SGBD: Sistema Gestor de Base de Datos. Es un conjunto de programas que permiten crear y mantener una Base de datos.

SIGOB: Sistema de Información de Gobierno.

Software: es el equipamiento lógico o soporte lógico de una computadora digital; comprende el conjunto de los componentes lógicos necesarios que hacen posible la realización de tareas específicas, en contraposición a los componentes físicos, que son llamados hardware.