Universidad de las Ciencias Informáticas

Facultad 6

Título: Sistema de Información de Gobierno. Mercado de datos Control de energía

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

Autores: Wendy Romalde Ruiz

Yezenia Verdecia Serrano

Tutora: Ing. Marisel Santana Rodríguez

Co-tutora: Ing. Mabel Medina Rodríguez

La Habana, Junio de 2011

"Año 53 de la Revolución"

DECLARACIÓN DE AUTORÍA

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

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

Wendy Romalde Ruiz

_____

Firma del Autor

Yezenia Verdecia Serrano

_____

Firma del Autor

Ing. Marisel Santana Rodríguez

_____

Firma de la Tutora

DATOS DE CONTACTO

Ing. Marisel Santana Rodríguez.

Categoría Docente: Instructor recién graduado

Graduado en el 2010 de Ingeniero en Ciencias Informáticas de la Universidad de Ciencias Informáticas (UCI).

e-mail: msantana@uci.cu

Ing. Mabel Medina Rodríguez.

Categoría Docente: Instructor

Graduado en el 2007 de Ingeniero en Ciencias Informáticas de la Universidad de Ciencias Informáticas (UCI).

e-mail: mmedina@uci.cu

AGRADECIMIENTOS

A mi madre que siempre ha sido mi ejemplo a seguir y a quien debo haberme convertido en la persona que soy.

A mi abuela Herminia, por todo su cariño y por enseñarme cada día a ser una mejor persona.

A mi novio Yosbel, que es mi vida y a quien amo con toda mi alma por ser la persona tan maravillosa que es.

A mi padre, por su preocupación constante y por cuidarme durante tantos años a pesar de todo.

A mi hermano Daniel, por todo lo que hemos compartido y por seguir siendo la bala perdida, simpático simpatiquísimo.

A Mirna y a Mili, que son como otra familia para mí y que siempre tendrán un lugar especial en mi corazón.

A mi gente de Villa Clara, especialmente a mi bisabuela Julia que es un ejemplo de fortaleza y perseverancia para toda su familia.

A mi suegrita Hortensia y al tía Angelina, por su amabilidad y forma de ser tan especial.

A mi familia de Miramar que siempre se preocupa por mi bienestar, especialmente mi abuela Irma.

A mis tutoras Mabel y Marisel, por todo lo que me han enseñado durante el desarrollo de esta tesis.

A toda la gente del departamento de Almacenes de datos por su colaboración para que esta tesis pudiera llevarse a cabo.

A Manuel y a Reina, por su apoyo y preocupación a pesar del poco tiempo que llevamos de conocernos.

A mis grandes amigas Ismaray, Viviana, Yudalmis y Maybe, por haberme aguantado estos cinco años sin quejarse.

A Yezenia mi compañera de tesis, porque a pesar de todas las dificultades que se nos presentaron en el camino pudimos terminar con éxito este trabajo.

A mis abuelos Rolando y Marcelino, que aunque ya no están presentes sé que siempre me están cuidando en donde quiera que estén.

Al Comandante Fidel, por haber hecho realidad esta universidad y por darme la oportunidad de estudiar en esta gran escuela.

Wendy

AGRADECIMIENTOS

A mis padres Marilin y Abel por estar en el momento en que más los necesito y brindarme su apoyo incondicional. Por todo el amor y cariño que constantemente me regalaron, por todo el esfuerzo que hicieron para lograr que me hiciera una profesional, sin ustedes no hubiera sido posible el éxito de mi carrera.

A mi hermano Abel Adrian por siempre darme tanto apoyo, cariño y por ser uno de los tesoros más preciados para mí, te quiero mucho mi hermanito.

A mi novio Yasel por ayudarme en todo los años de mi carrera y por tener que aguantar todas mis locuras, siempre que salía mal en una prueba. Gracias por apoyarme y ser el sostén de mi vida al estar lejos de mis padres, te amo mi amor.

A mi suegra Ivón, por ser en estos años de universidad como una madre para mí, por su constante preocupación y por poder contar con su apoyo siempre que lo necesite. Gracias por estar presente.

A toda mi familia, en especial a mis abuelos, Marina y Ernan, a mis tios Ernel, Macy Maricelis, y a mis primos Yurita y Yony.

A dos personas que desde que los conocí fueron como unos abuelos para mi, Roberto y Deysi. Gracias por su apoyo en estos años.

A mis amistades que han compartido conmigo durante estos cinco años, en especial Alejandro, Aylenis, Elisa, Leydi, Bety.

A mi compañera de tesis Wendy por ser una buena compañera durante todo el trabajo que hicimos.

A mi mis tutoras Marisel y Mabel por brindarme su experiencia profesional y ayudarme en todo lo que estuvo a su alcance.

A la Revolución y a nuestro comandante Fidel por darnos la oportunidad de estudiar en esta universidad de excelencia.

Gracias a todos aquellos que no están aquí, pero que me ayudaron a que este gran esfuerzo se volviera realidad.

Yezenia

DEDICATORIA

A mi familia por toda su preocupación y cariño y porque sin ella no hubiera podido llegar a ser quien soy.

A mi novio por su apoyo incondicional y por quererme tal como soy.

A todos aquellos que aunque no estén conmigo han contribuido a que este sueño se haga realidad.

Wendy

A mi mamá por saber comprenderme y darme su apoyo siempre que lo he necesitado. Gracias mami por todo el amor que me das y por todos tus cuidados en todo el tiempo que hemos vividos juntas.

A mi papá por ser exigente cuando tiene que serlo y por depositar toda su confianza en mí.

A mi hermano por ser el mejor hermano que una gran hermana puede tener, gracias mi pombito lindo.

Yezenia

RESUMEN

La constante evolución de las Tecnologías de la Información y las Comunicaciones (TIC), así como el incremento del uso de las mismas en la sociedad, ha provocado la búsqueda de nuevas y mejores soluciones que permitan resolver una amplia gama de problemas. Entre las principales problemáticas, a las que se hace necesario prestar especial atención, se encuentra el tema relacionado con la energía. La Oficina Nacional de Estadísticas (ONE), tiene como una de sus tareas analizar el comportamiento de los indicadores energéticos por los cuales se mide este recurso, para de esta forma apoyar la toma de decisiones respecto a su obtención y utilización en el país. Pero este proceso se realiza de forma manual por los especialistas de la ONE, lo que trae como consecuencia que el análisis del inmenso volumen de información que se recoge, sea un trabajo complejo, ya que no existe una forma de integrar los datos de manera automática. En este sentido, la Universidad de las Ciencias Informáticas de mutuo acuerdo con la Oficina Nacional de Estadísticas, ha decidido elaborar un mercado de datos que permita centralizar toda la información referente a la energía, para facilitar la integración de los datos, el análisis estadístico y apoyar la toma de decisiones a nivel nacional.

Palabras claves: decisiones, energía, estadística, indicadores energéticos

ÍNDICE

Introducción

El control de los datos relacionados con la energía en un país, constituye una tarea estratégica debido al impacto que puede tener en la economía de cualquier nación. El análisis estadístico ha jugado un papel fundamental en este sentido, ya que desde su surgimiento ha sido utilizado para apoyar la toma de decisiones, especialmente las relacionadas con el gobierno y los órganos administrativos, aunque también son ampliamente empleadas en el mundo de los negocios debido a los beneficios que ofrece.

En Cuba, desde el triunfo de la Revolución, la política energética ha estado orientada a satisfacer las necesidades del pueblo cubano, para lo cual se hizo necesaria la toma de medidas que garantizaran un mejor control sobre los recursos energéticos. Tal situación, gana mayor relevancia con el inicio de la revolución energética llevada a cabo desde hace algunos años, con vistas a promover el ahorro de energía en todo el país.

La Oficina Nacional de Estadísticas (ONE), órgano rector de la estadística en Cuba, tiene como misión garantizar la obtención de estadísticas de calidad a través del Sistema de Información Estadístico Nacional (SIEN), así como su adecuada difusión de acuerdo con las necesidades del país en información estadística. Uno de los aspectos más importantes en los que trabaja esta organización, es el control de la obtención y el uso de los portadores energéticos en Cuba. La información referente a este tema, es recogida a través de diferentes modelos en cada entidad del territorio nacional, los cuales se envían a las sedes municipales de la ONE correspondientes y posteriormente a las sedes provinciales. Finalmente, son digitalizados en ficheros con formato ´dbf´ y enviados a la oficina central situada en la capital cubana.

Desde su surgimiento, la Universidad de las Ciencias Informáticas (UCI) se ha desempeñado no sólo como una institución educacional, sino también como un centro productor de software con proyectos de carácter nacional e internacional. El Centro de Tecnologías de Gestión de Datos (DATEC) de la UCI, en conjunto con la Oficina Nacional de Estadísticas, decidió comenzar el proyecto Sistema de Información de Gobierno (SIGOB), para contribuir a la toma de decisiones en las diferentes áreas del SIEN.

El área donde se gestiona la información referente al control de la energía, es de gran importancia por la necesidad que existe en el país, de tomar decisiones que contribuyan al ahorro energético en todos los sectores posibles. Durante el estudio de la situación existente en la ONE, se identificaron algunos elementos que afectan la gestión de la información de esta área, los cuales se exponen a continuación.

Primeramente, se detectó que el análisis de la información recibida se realiza de forma manual por los especialistas, tarea que resulta muy engorrosa debido principalmente a que los datos se encuentran dispersos en múltiples ficheros, cantidad que se va incrementando con el paso del tiempo. Unido a esto, la existencia de diferentes fuentes de procedencia, trae como consecuencia la aparición de inconsistencias en los datos, influyendo negativamente en la calidad de los resultados obtenidos.

Dicha situación, conlleva a que existan limitaciones para recuperar indicadores desde distintas perspectivas de análisis, además de provocar que el proceso de recuperación y elaboración de informes resulte costoso en esfuerzo y tiempo, factores que dificultan la disponibilidad de información estadística para los altos mandos del estado cubano.

Debido a la situación anteriormente descrita, se establece como **problema de la investigación**: ¿Cómo contribuir a la toma de decisiones en el área Control de energía del Sistema de Información de Gobierno?

Dicho problema, conlleva a definir como **objeto de estudio** de la presente investigación: los almacenes de datos, delimitando el **campo de acción** al mercado de datos para el área Control de energía del Sistema de Información de Gobierno.

Para dar solución al problema de la investigación, se identifica como **objetivo general**: desarrollar el mercado de datos Control de energía del Sistema de Información de Gobierno que contribuya a la toma de decisiones. En correspondencia con el mismo, se plantean los siguientes **objetivos específicos**:

1. Realizar el análisis y diseño del mercado de datos.
2. Implementar el mercado de datos.
3. Validar el mercado de datos.

Para darle cumplimiento a los objetivos planteados, se proponen las siguientes **tareas de la investigación**:

1. Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
2. Modelación del negocio.
3. Levantamiento de requisitos.
4. Diseño del modelo de datos.
5. Diseño del subsistema de integración.
6. Diseño del subsistema de visualización.

7. Diseño de los casos de pruebas.
8. Implementación del subsistema de integración.
9. Implementación del subsistema de visualización.
10. Aplicación de las listas de chequeo.
11. Aplicación de las pruebas de aceptación.
12. Aplicación de los casos de pruebas.

Estructura de la Tesis

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

Capítulo 1: Fundamentos teóricos sobre el desarrollo de un mercado de datos

En el capítulo, se abordan los principales conceptos relacionados con los Almacenes de datos, sus principales características y las ventajas que proporciona su uso en el mundo empresarial. Se exponen además, aspectos fundamentales tales como la metodología de desarrollo a utilizar y las herramientas más empleadas en los procesos de integración de datos e inteligencia de negocio a nivel internacional.

Capítulo 2: Análisis y diseño de un mercado de datos

En el capítulo, se realiza un análisis del negocio con el propósito de comprender mejor los aspectos de mayor relevancia y se propone un diseño de la solución, con las características necesarias para satisfacer las necesidades manifestadas por el cliente.

Capítulo 3: Implementación del mercado de datos

El capítulo está dirigido a la implementación de los diferentes aspectos relacionados con los procesos de integración de datos, con el propósito de brindar una mayor comprensión de las estrategias y procedimientos utilizados. Además, se abordan elementos relacionados con la implementación de la capa de inteligencia de negocio, incluyendo la creación de las estructuras necesarias para la navegación y el análisis de los datos.

Capítulo 4: Pruebas

En el capítulo se exponen las pruebas realizadas al mercado de datos, así como los resultados obtenidos en cada una de ellas luego de su aplicación. Dichas pruebas son realizadas para garantizar el cumplimiento de las exigencias del cliente y la calidad del producto final.

Capítulo 1: Fundamentos teóricos sobre el desarrollo de un mercado de datos

1.1. Introducción

En el capítulo, se abordan los principales conceptos relacionados con los Almacenes de datos, sus principales características y las ventajas que proporciona su uso en el mundo empresarial. Se exponen además, aspectos fundamentales tales como la metodología de desarrollo a utilizar y las herramientas más empleadas en los procesos de integración de datos e inteligencia de negocio a nivel internacional.

1.2. Almacenes de Datos

Para cualquier empresa, resulta de gran interés analizar e interpretar la información que genera, con el fin de contar con elementos que le permitan tomar decisiones para maximizar sus ganancias o mejorar su rendimiento. Con el desarrollo de las tecnologías, el volumen de información generada fue aumentando considerablemente, dando lugar a grandes cantidades de datos históricos, cuyo análisis resultaba cada vez más complejo.

Los Almacenes de Datos (AD), en inglés Data Warehouse (DWH), surgen como una solución a esta problemática, con el fin de apoyar el proceso de toma de decisiones en las empresas. La definición universalmente aceptada de un DWH es la expresada por William H. Inmon, quien plantea que:"…*Un Almacén de Datos consiste en una colección de datos orientada al negocio, integrada, no volátil y variante en el tiempo, para el apoyo a la toma de decisiones administrativas*." [1]

Por su parte, Ralph Kimball quien es una de las personalidades más influyentes en el área, propone otra definición al catalogarlo como: "…*una copia de datos transaccionales, específicamente estructurados para la consulta y el análisis*". [2]

A pesar de plantear conceptos un poco diferentes, los dos convergen en un mismo punto y es que el DWH tiene como tarea fundamental, organizar la información recogida de diferentes fuentes para facilitar su análisis y apoyar el proceso de toma de decisiones.

Luego de un estudio de la bibliografía existente sobre el tema, se puede concluir que un almacén de datos es una colección de datos, que sirve de apoyo para la toma de decisiones administrativas y que debe cumplir las siguientes características:

Orientada: Debe estar orientado hacia los aspectos que resulten de interés para la organización. Por consiguiente, la información es organizada por temas garantizando un acceso rápido y fácil a la misma.

Integrada: Un almacén se nutre de diversas fuentes heterogéneas, por lo que resulta muy común encontrar la misma información almacenada de diferentes formas. Por tal motivo, los datos antes de ser cargados, deben pasar por un proceso de integración donde se corrigen, validan y estandarizan, para lograr llevarlos a un formato único que facilite su consulta en el DWH.

Variable en el tiempo: Los cambios producidos con el paso del tiempo en los datos almacenados también se van registrando en el almacén, lo que permite brindar una perspectiva histórica de la información, que abarca un espacio de tiempo de cinco a diez años.

No volátil: Los datos recogidos en el almacén van a poder ser leídos por los usuarios, pero no podrán ser modificados. Solamente se podrán introducir nuevos datos en el almacén o consultar los que ya se encuentran almacenados.

1.1.2. Componentes de un almacén de datos

Un DWH está compuesto por un conjunto de elementos, necesarios para lograr el cumplimiento de sus objetivos. A continuación, se ofrece una explicación de cada uno de ellos [2]:

Sistemas Fuentes Operacionales: son los sistemas utilizados en las empresas para gestionar sus transacciones, información que es almacenada en diferentes formatos de acuerdo a las necesidades del negocio. Estos sistemas, conservan pocos datos históricos, pues generalmente realizan salvas de la información para trabajar con los datos generados en un corto período de tiempo y de esta forma hacer las recuperaciones más fácilmente. Las prioridades principales que poseen son el procesamiento del rendimiento y la disponibilidad.

Área de procesamiento (staging area): es un área de almacenamiento donde se realizan un conjunto de procesos comúnmente conocidos como extracción, transformación y carga (ETL), en los cuales se invierte la mayor cantidad de tiempo y esfuerzo durante la construcción de un DWH. Primeramente, se realiza la extracción de los datos necesarios para el almacén de las diferentes fuentes, para luego pasar por un proceso de transformación donde se eliminan errores e inconsistencias que dificulten su posterior análisis. Finalmente, una vez que los datos están listos para ser almacenados, son cargados en el área de presentación del DWH.

Área de presentación: en esta área los datos son almacenados, organizados y puestos a disposición de los usuarios para ser consultados, analizados o realizar reportes sobre ellos. En ella, se almacena toda la

información que puede ser de utilidad para el proceso de toma de decisiones en la empresa, diseñada mediante esquemas dimensionales. Generalmente, es referenciada como un conjunto de mercados de datos integrados, donde cada uno representa a un proceso específico del negocio.

Herramientas de acceso a los datos: en este componente, se utiliza la palabra herramienta para referirse a la variedad de habilidades que pueden ser provistas a los usuarios del negocio, para soportar el proceso de toma de decisiones. Por definición, su actividad fundamental consiste en consultar la información que se encuentran en el área de presentación, lo que constituye el objetivo principal de los almacenes de datos.

Figura 1: Componentes de un *Data Warehouse*

1.1.3. Ventajas y desventajas de los almacenes de datos

La utilización de un almacén de datos en una empresa, provee numerosos beneficios en cuanto al proceso de toma de decisiones administrativas, las cuales influyen fuertemente en su desarrollo. Las facilidades que brinda para consultar y analizar información histórica, resulta muy provechoso en el mundo empresarial actual, donde constantemente se necesitan nuevas estrategias que se adapten a los cambios tan frecuentes que sufre el mercado.

Pero también presenta sus desventajas, las cuales pueden ser un impedimento para la utilización de este tipo de sistema en una entidad determinada. A continuación, se exponen los argumentos necesarios que expresan de forma más clara lo explicado anteriormente:

Beneficios

- Transforma datos orientados a las aplicaciones, en información orientada a la toma de decisiones.
- Permite un análisis inmediato de las actividades de la empresa.
- Agilidad en el control de los activos tangibles de la empresa.
- Capacidad de analizar y explorar las diferentes áreas de trabajo.
- Relación total con el cliente.
- Facilidades en la gestión y análisis de recursos.
- Permite establecer una conexión entre los departamentos empresariales que son independientes y el resto.
- Reacciona rápidamente a los cambios del mercado.

Inconvenientes

- Gran inversión que supone este tipo de proyectos.
- La tecnología no se encuentra del todo madura.

Los almacenes de datos, han llegado a convertirse en una potente herramienta para el mundo empresarial, donde cada vez resulta más imperioso contar con mecanismos que brinden la información necesaria para aumentar la productividad, mejorar las ventas o detectar los fraudes que se cometen en las empresas. Un sistema que trabaje en correspondencia con las necesidades y objetivos reales de cualquier entidad, puede ser un instrumento muy valioso, por lo que debe ser utilizado de manera eficiente para poder obtener los mejores resultados.

1.3. Mercados de Datos

Muchos son los autores que han expresado su opinión en cuanto a la definición del término mercado de datos, en inglés Data Mart (DM), pero una de las más abarcadoras es la propuesta por Ralph Kimball quien considera que constituye "*…un conjunto flexible de datos, idealmente basado en el dato más atómico posible (granular) para ser extraído de las fuentes operacionales y presentado en un modelo simétrico (dimensional),*

que es más resistente cuando se enfrentan con las más inesperadas consultas de los usuarios (…). Podemos decir que los data marts están conectados con la arquitectura de los DWH en su forma más simple y que representan los datos de un sólo proceso del negocio a la vez.".[2]

Tomando como referencia lo planteado por Ralph Kimball, se puede afirmar que un DM se especializa en el almacenamiento de los datos de un área específica del negocio, a diferencia de un DWH que gestiona la información a nivel de empresa. Un DM, se caracteriza por estructurar la información de forma tal que se alcance el más alto nivel de detalle, condición que facilita su análisis desde todas las perspectivas posibles que puedan afectar al área del negocio que le corresponda. Al igual que un almacén de datos, puede ser una alternativa de solución a tener en cuenta para resolver los problemas planteados anteriormente, ya que son similares en cuanto a diseño y construcción.

Debido a que existen similitudes entre los DWH y los DM, algunas personas tienden a pensar que se trata de lo mismo. En la siguiente tabla, se muestran algunos elementos que demuestran lo contrario. [3]

Almacén de datos	**Mercado de Datos**
• Corporativo o red empresarial. • Está conformado por la unión de todos los DM. • Los datos son recibidos desde el área de procesamiento. • Consultas sobre la presentación de recursos. • Estructura para vista corporativa de los datos.	• Departamental. • Constituye un simple proceso del negocio. • Unión en forma de estrella (hechos y dimensiones). • Tecnología óptima para el acceso a los datos y el análisis. • Estructura para adaptarse a la vista de los datos departamentales.

Tabla 1: Comparación entre DWH y DM

1.4. Integración de datos

Uno de los mayores desafíos que se enfrentan durante el desarrollo de un DWH, es el problema de transformar los datos obtenidos de las diferentes fuentes de información, de manera que sea más fácil manipularlos una vez cargados en el almacén. En cada una de estas fuentes, los datos son almacenados de manera que sean entendidos por el sistema que va a utilizarlos, por lo que existen numerosas formas de almacenar la misma información.

Esta situación trae como consecuencia que no exista uniformidad en los datos, lo que unido a las grandes cantidades de información que se generan en una empresa, provoca que el trabajo con ellos sea aún más complejo.

Por tal motivo, se hace necesario realizar la integración de los datos, ya que mediante este proceso se eliminan muchos errores e inconsistencias que pueden afectar la calidad de los reportes generados posteriormente por los especialistas, dificultando la toma de decisiones en la organización.

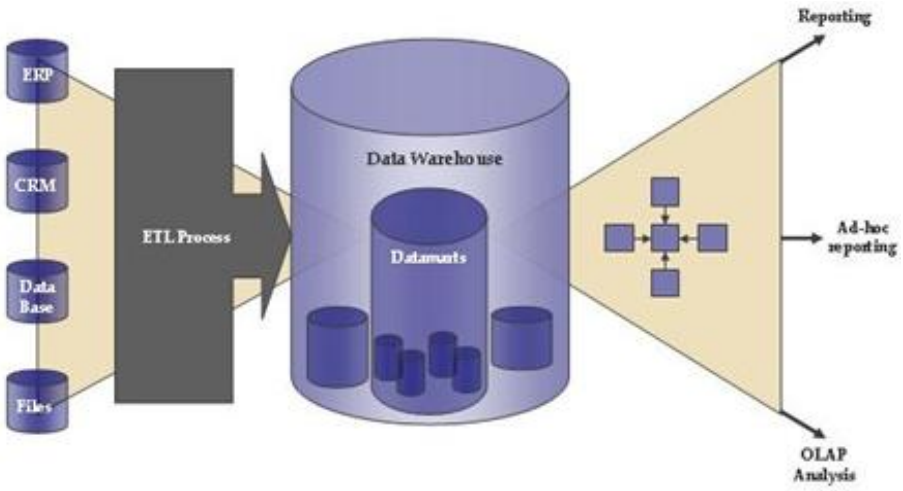


Figura 2: Proceso de integración de datos

La integración de datos, consiste en el proceso de unificación de los datos provenientes de múltiples fuentes y puede realizarse de diferentes formas, ya sea mediante la replicación de datos, la integración de información empresarial o a través de los procesos de ETL. Esta última variante, es la utilizada actualmente por el centro DATEC y la seleccionada para aplicar en el proyecto SIGOB debido a ciertas características que posee, algunas de las cuales son brevemente explicadas a continuación.

Características del proceso de ETL

- Es un mecanismo de carga muy eficiente y efectivo orientado a los DWH.
- Enfocado a migrar y mezclar datos.
- Reduce la exposición a desarrollos manuales (codificación), producto de la existencia en el mercado de herramientas potenciales para la implementación visual, con manejo de excepciones, gestión y planificación de tareas.
- Necesita pocos servicios de administración y mantenimiento.
- Gran capacidad para llevar a cabo transformaciones.

1.5. Inteligencia de negocios

Las soluciones de inteligencia de negocios, en inglés Business Intelligence (BI), permiten a las empresas contar con una vía rápida y fácil que les ayude en el proceso de toma de decisiones. Muchos son los conceptos que se han planteado por diferentes especialistas relacionados con este término, algunos de los cuales se muestran a continuación:

"*no es una metodología, software, sistema o herramienta específica, es más bien una colección de tecnologías que van desde arquitecturas para almacenar datos, metodologías, técnicas para analizar información y software, entre otros, con un fin común para el apoyo a la toma de decisiones*". [5]

"…*Son los procesos, tecnologías y herramientas que se necesitan para convertir los datos en información, la información en conocimiento y el conocimiento en planes que impulsan acciones rentables para el negocio. La Inteligencia de Negocios abarca el almacenamiento de datos, herramientas analíticas, y contenido y gestión del conocimiento*…" [6].

"…*conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada en información estructurada, para su explotación directa o para su análisis y conversión en conocimiento que soporte la toma de decisiones sobre el negocio*".[7]

Luego de analizar los conceptos antes planteados, se puede definir la inteligencia del negocio como el proceso de analizar los datos de una empresa para extraer conocimientos que respalden las decisiones empresariales. Además, el término BI también es considerado como una colección de conceptos, lo cual le

da un poder enorme, ya que pueden integrarse funciones que tradicionalmente estaban separadas, tales como el acceso a datos, generación de reportes, explotación, pronóstico y análisis.

Las soluciones de BI dentro de empresas muy grandes, se han convertido en un apoyo indispensable para la toma de decisiones en cualquier nivel de la organización, pues proporcionan amplias ventajas ya sean competitivas para las empresas o de índole cognoscitivo en temas no empresariales.

Entre las principales ventajas que ofrece una solución de BI se encuentran las siguientes: [8]

- Control de costes, al tener un solo sistema que permita manejar fácilmente los distintos programas que se encuentran en los diferentes departamentos de una compañía.
- Mejora de la colaboración y la calidad de las decisiones, facilitando el acceso a la información en todos los niveles de la organización.
- Orienta las soluciones tecnológicas hacia el usuario, ya que reduce los tiempos de aprendizaje mediante el manejo de herramientas de uso cotidiano.
- Proporciona una profunda visión del negocio a través de sistemas integrados como cuadros de mando integrales, tableros digitales, consultas y reportes, minería de datos y almacenamiento analítico.
- Asiste a los ejecutivos para planear y pronosticar el trabajo, presentando una descripción común de los procesos del negocio de una compañía.

Propósitos de las soluciones de inteligencia de negocio

- Convertir grandes volúmenes de datos, en un valor para el negocio a través de los reportes analíticos.
- Generar información para el control de los procesos del negocio, independientemente de la fuente de datos.
- Soportar la toma de decisiones.
- Diferenciar la información útil para los usuarios finales.
- Uniformizar los términos usados en la institución, independientemente del origen de los datos o de la forma de extracción, transformación y agregación.

1.5.1. Técnicas de Almacenamiento de Datos

El Procesamiento Analítico en Línea (OLAP por sus siglas en inglés), es una solución que se utiliza en la inteligencia de negocio, cuyo principal objetivo se basa en lograr que el proceso de consultas de grandes cantidades de datos sea más rápido. Para ello, utiliza estructuras multidimensionales o cubos OLAP, que

contienen datos resumidos de grandes bases de datos. De manera general, el procesamiento analítico en línea puede verse como la síntesis, análisis y consolidación de grandes volúmenes de datos con un enfoque multidimensional.

La razón de usar OLAP para las consultas, es la velocidad de respuesta que proporciona a los usuarios. Una base de datos relacional, almacena entidades en tablas discretas si han sido normalizadas, estructura que es beneficiosa en un sistema OLTP (Procesamiento de Transacciones en Línea), pero cuando se realizan consultas a varias tablas de forma simultánea, resulta muy compleja y lenta. Por esta razón, se necesita una base de datos multidimensional debido a que son mejores para realizar operaciones de búsqueda.

Existen tres modelos para realizar este proceso: ROLAP, MOLAP y HOLAP. Seguidamente, se explican brevemente las características principales de cada uno de ellos.

ROLAP

En el Procesamiento Analítico Relacional en Línea, en inglés *Relational Online Analytical Process* (ROLAP), los datos son guardados en filas y columnas de forma racional. Este modelo presenta los datos a los usuarios en forma de dimensiones del negocio, con el objetivo de ocultar las estructuras de almacenamiento a los usuarios y presentar los datos multidimensionalmente.

El modelo ROLAP, es usado fundamentalmente sobre la información que no se consulta frecuentemente, debido a que resulta muy útil cuando se desea consultar información que se almacena durante muchos años. En la siguiente figura se muestra la forma de almacenamiento de ROLAP:

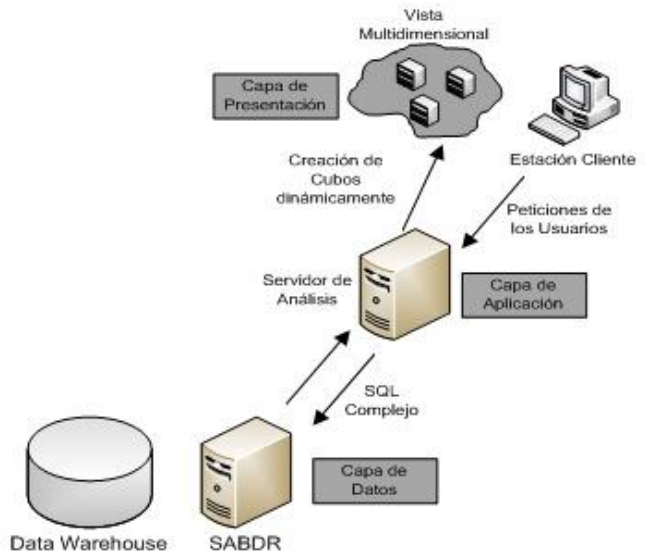


Figura 3: Modelo de almacenamiento ROLAP

MOLAP

El Procesamiento Analítico Multidimensional en Línea, en inglés *Multidimensional Online Analytical Process* (MOLAP), almacena los datos multidimensionalmente a diferencia del modelo ROLAP. La estructura de los datos en este modelo, es estática para que la lógica al procesar el análisis multidimensional pueda ser basada en métodos bien definidos, con el propósito de establecer las coordenadas del almacenamiento de los datos.

Para realizar el acceso a la información almacenada de forma más rápida y efectiva, las estructuras de almacenamiento se organizan en grandes arreglos dimensionales, que son una copia de la fuente de datos y persisten físicamente en la misma estación de trabajo donde está instalado el DWH. En la siguiente figura se muestra la forma de almacenamiento de MOLAP.

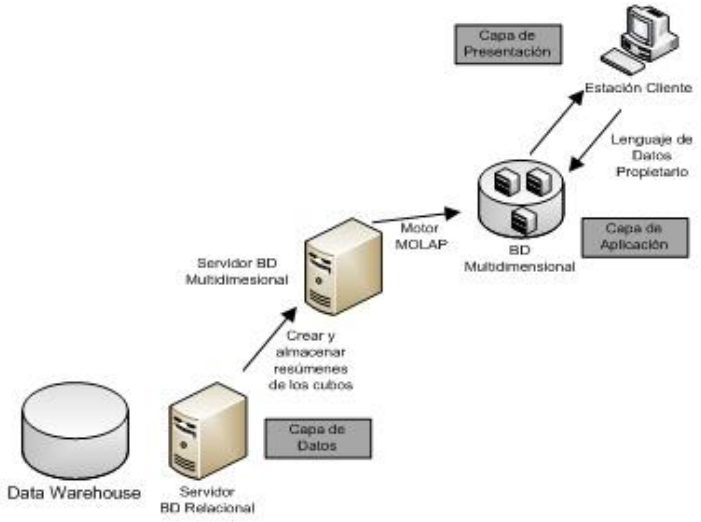


Figura 4: Modelo de almacenamiento MOLAP

HOLAP

El modo de almacenamiento Procesamiento Analítico Híbrido en Línea, en inglés *Hybrid Online Analytical Process* (HOLAP), como su nombre lo indica, es un híbrido entre los métodos ROLAP y MOLAP, que permite almacenar una parte de los datos como en un sistema MOLAP y el resto como en uno ROLAP. Además posee dos tipos de particionamiento:

Particionamiento vertical: Almacena las agregaciones como un MOLAP para mejorar la velocidad de las consultas, y los datos se detallan en ROLAP para optimizar el tiempo en que se procesa el cubo.

Particionamiento horizontal: En este modo, HOLAP almacena una sección de los datos. Normalmente, los datos más recientes se almacenan en modo MOLAP para mejorar la velocidad de las consultas y los datos más antiguos se guardan en ROLAP.

Comparación entre ROLAP y MOLAP

Cuando se debe escoger entre un modelo y otro, es importante analizar el rendimiento de las consultas para los usuarios y tener en cuenta las tecnologías disponibles que se van a utilizar en los diferentes modelos. En la siguiente tabla, se muestra una comparación entre los modelos ROLAP y MOLAP, basándose en el almacenamiento de los datos, tecnologías subyacentes, funciones y características. [3]

Modelo	Almacenamiento de Datos	Tecnologías Subyacentes	Funciones y Características
ROLAP	Almacenamiento como tablas relacionales	Uso de SQL complejo para obtener los datos del depósito	Ambiente conocido y disponibilidad de herramientas
	Resumen detallado de los datos disponibles	Motor ROLAP en el servidor de análisis, crea los cubos de datos sobre la marcha	Limitación en funciones de análisis complejas
	Volúmenes altos de datos	Vistas multidimensionales en la capa de presentación	Realización de agregaciones no siempre son fáciles
	Todos los datos de acceso están en la bodega almacenados		
MOLAP	Almacenamiento como tablas relacionales	Creación de cubos de datos prefabricados por el motor MOLAP	Acceso rápido
	Diversos resúmenes de datos se mantienen en las BD propietarias	Tecnología propietaria para almacenar las vistas multidimensionales en arreglos, no en tablas	Fuerte librería de funciones para cálculos complejos

	Volúmenes de datos moderados	Matriz de alta velocidad para la recuperación de los datos	Facilita el análisis independientemente de la cantidad de dimensiones
	Resúmenes de acceso a datos detallados en BD multidimensionales	Escasa tecnología de matriz de datos para gestionar la escasez de los resúmenes	Amplitud en la capacidad de acción en el Drill-Down y Slicing-Dicing

Tabla 2: Comparación entre ROLAP y MOLAP

Tomando como base los elementos expuestos en la tabla de comparación, se decide seleccionar para el desarrollo de la presente investigación el modelo ROLAP.

1.6. Metodología de desarrollo de Almacenes de Datos

Crear un producto con calidad y coste reducido, es uno de los retos más importantes que se imponen los desarrolladores de software, lo cual depende en gran medida de que las herramientas y procesos que se seleccionen sean los más adecuados. Con el propósito de mejorar estos indicadores se crearon diferentes metodologías de desarrollo, las cuales definen un conjunto de pasos y procesos a seguir, que permiten estructurar, controlar y planificar el proceso de desarrollo de software.

El diseño de un almacén de datos, es una disciplina que ha ido madurando con el paso de los años, tiempo durante el cual, han sobresalido dos enfoques principales concernientes a dos de las personalidades más influyentes en el área, Ralph Kimball y William H. Inmon. La principal diferencia entre ambos estilos radica en la forma de enfrentar el problema.

La metodología propuesta por Kimball consiste en un método iterativo, pues plantea que el almacén se debe construir pieza por pieza, es decir, un DM a la vez. De esta forma, concibe la construcción del DWH mediante la unión de todos los DM que se desarrollen, solución que es muy eficaz y conduce a una solución completa en un período de tiempo muy corto. Debido a sus características, esta metodología es reconocida como bottom-up por su arquitectura ascendente y como existe abundante documentación sobre ella, resulta una buena opción para aplicar durante el desarrollo de un almacén de datos.

Sin embargo, el enfoque propuesto por Inmon presenta una arquitectura descendente por lo que se conoce como top-down. Este método, propone la creación de un repositorio de datos corporativo como fuente de

información consolidada, persistente, histórica y de calidad, donde la información se encuentre en 3FN. De esta forma, se crea primero un DWH del cual se van a nutrir los diferentes DM que van a satisfacer las necesidades específicas de cada proceso del negocio.

1.6.1. Justificación de la metodología a utilizar

Para tratar de solucionar las dificultades identificadas en el área de energía del SIEN, se decide utilizar la metodología utilizada en el centro para el desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio (DW&BI), la cual tiene como base el enfoque planteado por Kimball, pero se adapta a las características particulares de la UCI y del centro. Los elementos que influyeron en esta decisión, se exponen a continuación:

- Crea los conceptos de hechos y dimensiones, lo que indudablemente es muy eficaz en el proceso de la toma de decisiones y proporciona mayor agilidad en el proceso de desarrollo.
- Propone ir construyendo el almacén de datos a través de la construcción de los mercados de datos departamentales, lo que constituye una buena estrategia y coincide con la división lógica de las empresas, entidades y organismos.
- Existe abundante documentación sobre la misma, la respuesta a todas las dudas y preguntas que puedan surgir se pueden encontrar en la web, a través de los servicios que brinda el grupo creador de la metodología.
- Es una metodología madura y reconocida por el resto de la comunidad dedicada al tema. Tiene bien definidas las etapas, actividades, artefactos y roles.
- Como complemento a la misma y fortaleciendo la etapa del levantamiento de requisitos; se toma lo planteado por Leopoldo Zenaido Zepeda Sánchez en su Tesis de Doctorado, orientando así el trabajo a los casos de uso, de forma que se logre estar más alineado con las tendencias y normas de la Universidad.[9]

1.7. Herramientas de modelado

Las herramientas CASE (*Computer Aided Software Engineering*), son un conjunto de métodos, utilidades y técnicas, que facilitan la automatización del ciclo de vida del desarrollo de un software. Para el modelado de la presente investigación, se decide utilizar Visual Paradigm for UML en su versión 6.4, debido a que es una de las herramientas UML más utilizadas para realizar el ciclo de vida completo del desarrollo de un software:

análisis, diseño, implementación y despliegue. Es muy fácil de usar, ya que posee una interfaz gráfica amigable y posibilita diseñar todos los tipos de diagramas de clases necesarios, generar código desde cualquiera de estos diagramas y generar documentación. Además, es fácil de instalar y actualizar y sus diferentes ediciones son compatibles entre ellas. [10]

1.8. Sistema Gestor de Bases de Datos

Se define como un conjunto de programas que administran y gestionan la información contenida en una base de datos, contribuyendo en la realización de las siguientes acciones: [11]

- Definición de los datos
- Mantenimiento de la integridad de los datos dentro de la base de datos
- Control de la seguridad y privacidad de los datos
- Manipulación de los datos

PostgreSQL v8.4: es un potente motor de bases de datos, que tiene prestaciones y funcionalidades equivalentes a muchos gestores de bases de datos comerciales. Está considerado como el gestor de bases de datos de código abierto más avanzado del mundo, ya que proporciona un gran número de características que normalmente sólo se encontraban en gestores comerciales como DB2 u Oracle. [12]

A continuación se listan las características que más lo identifican y que justifican su utilización en el proyecto SIGOB:

- PostgreSQL aproxima los datos a un modelo objeto-relacional, siendo capaz de manejar complejas rutinas y reglas, como por ejemplo: consultas SQL declarativas, control de concurrencia multi-versión, soporte multi-usuario, optimización de consultas y herencia.
- Es altamente extensible pues soporta operadores, funciones, métodos de acceso y tipos de datos definidos por el usuario.
- Soporta la integridad referencial, la cual es utilizada para garantizar la validez de los datos en una base de datos.
- Posee un API flexible.

1.9. Herramienta para el perfilado de los datos

El perfilado de los datos, es una de las primeras tareas a realizar en el proceso de calidad de datos y consiste en realizar un análisis inicial sobre los datos de las fuentes, con el propósito de empezar a conocer su estructura, formato y nivel de calidad [13]. Una de las ventajas que ofrece, es que consigue interactuar con archivos en formato XML y ficheros creados en Microsoft Excel, lo que resulta de gran utilidad ya que gran parte de las fuentes de datos son compatibles con este tipo de formato.

Además, posibilita generar informes y gráficos que permiten a los usuarios determinar rápidamente el nivel de calidad de los datos. También, resulta de gran ayuda cuando se desean combinar resultados y gráficos, pues permite crear vistas fáciles de interpretar para evaluar la calidad de la información.

Debido a las ventajas anteriores y por ser la herramienta manejada en el Centro de Tecnologías de Gestión de Datos de la UCI, se decide utilizar el DataCleaner v1.5.3 para realizar el perfilado de los datos relacionados con el control de la energía. Igualmente, al tratarse de una aplicación de código abierto y licencia de uso gratuita, su uso implica un gran ahorro de inversión para el proyecto.

1.10. Herramientas de desarrollo

En la actualidad, existe una amplia variedad de herramientas de código abierto para realizar la integración de datos. Ejemplo de ellas, se pueden mencionar el *Scriptella Open Source ETL tool*, el *Talend Open Studio* y el *Pentaho Data Integration*, también conocido como *Kettle*, siendo esta última una de las que más se destaca. Seguidamente, se abordarán algunos de los elementos que la caracterizan:

Pentaho Data Integration

Es una herramienta muy intuitiva y fácil de utilizar, que permite definir transformaciones de forma gráfica, interconectando bloques que tienen diversas funciones. Es tremendamente versátil, ya que posee bloques que facilitan la lectura y escritura de cualquier base de datos o ficheros con formato Excel, Access y otros. Estos componentes permiten operar con los campos renombrando, normalizando, realizando cálculos, mapeando valores, realizando búsquedas auxiliares en bases de datos y normalizando/desnormalizando los datos de distintas filas en una sola. Las transformaciones que se hacen con esta herramienta, se guardan en un fichero ´.ktr´ que luego puede ser ejecutado desde una línea de comandos o un fichero batch. Debido a estas funcionalidades, es ampliamente utilizada en los procesos de extracción, transformación y carga de los datos. [14]

Estos factores, determinan la selección del Pentaho Data Integration v4.1.0 como la herramienta a utilizar para el proceso de integración de datos, la cual además es la utilizada en el centro DATEC, debido a que se cuenta con una amplia experiencia por parte de los especialistas en cuanto al trabajo con ella.

1.11. Herramientas para la inteligencia de negocios

Las herramientas de inteligencia de negocio, han sido creadas para ayudar a la toma de decisiones entre las diferentes empresas e instituciones de un estado. Además, permiten mostrar una visión general de todos los procesos de la entidad a sus directivos, facilitando un mejor entendimiento en el análisis y en la presentación de los datos. A continuación se presentarán algunas de estas herramientas y sus principales características.

Pentaho Schema Workbench v3.2: es un entorno visual para el desarrollo y prueba de cubos OLAP, los cuales proveen un mecanismo para buscar datos con rapidez y tiempo de respuesta uniforme, independientemente de la cantidad de datos en el cubo o la complejidad del procedimiento de búsqueda. Con esta aplicación, se puede configurar una conexión ´JDBC´ como el modelo físico, para luego elaborar el esquema lógico de manera simple y efectiva. Para ello, el entorno ofrece un editor de esquemas con la fuente de datos subyacente para su validación, además de permitir la ejecución de consultas ´MDX´ contra el esquema y la base de datos. [15]

Mondrian OLAP Server v3.0.4: es un servidor OLAP de código abierto muy popular, que gestiona la comunicación entre una aplicación OLAP escrita en Java y la base de datos con los datos fuentes. El núcleo del servidor Mondrian es similar a ´JDBC´, pero exclusivo para OLAP. Este proporciona la conexión a la base de datos y ejecuta las sentencias ´SQL´. Entre sus principales características, se encuentra las facilidades que brinda para el análisis de grandes volúmenes de información, que se encuentren almacenados en bases de datos que soporten ´JDBC´. [16]

Para el desarrollo de la capa de inteligencia de negocio, se seleccionan estas herramientas ya que a pesar de ser las utilizadas en el centro, poseen las características necesarias para realizar esta última etapa en el desarrollo de un almacén de datos.

Apache Tomcat v5.5: fue creado bajo el proyecto Jakarta de la Fundación Apache, siendo mantenido por una comunidad de desarrollo, logrando destacar como un producto robusto y altamente eficiente. Esta herramienta es gratis, fácil de instalar y se puede ejecutar en máquinas con pocos recursos, así como también es compatible con las API más recientes de Java. Debido a que su código binario posee un tamaño

total de un poco más de un megabyte, no ocupa mucho espacio de modo que no resulta extraño que se ejecute tan rápidamente. Además, su remarcada estabilidad y la capacidad de ejecución multiplataforma, lo han colocado como uno de los servidores más utilizados para el despliegue de aplicaciones web basadas en la tecnología Java. [17]

1.12. Conclusiones del capítulo

En el presente capítulo se abordaron los principales conceptos relacionados con los DWH, así como sus principales características, ventajas y desventajas. Luego del estudio de la bibliografía sobre el tema, se seleccionó como metodología de desarrollo la utilizada por el centro, que tiene como base la propuesta por Ralph Kimball. Como herramienta para la integración de los datos se escogió el Pentaho Data Integration (PDI) y el DataCleaner para realizar el perfilado de datos. Para el proceso de inteligencia de negocio, se decidió utilizar el Pentaho Schema Workbench y el Mondrian OLAP Server. Además, se seleccionó como servidor web el Apache Tomcat y como gestor de base de datos PostgreSQL.

Capítulo 2: Análisis y diseño de un mercado de datos

2.1. Introducción

En el capítulo, se realiza un análisis del negocio con el propósito de comprender mejor los aspectos de mayor relevancia y se propone un diseño de la solución, con las características necesarias para satisfacer las necesidades manifestadas por el cliente.

2.2. Estudio preliminar del negocio

Los portadores energéticos son aquellos recursos materiales o artificiales, que debido a ciertas propiedades o características que poseen, pueden ser transformados o procesados con la finalidad de obtener energía. Mantener un control sobre estos recursos, resulta de gran importancia debido al impacto positivo o negativo que puede tener la utilización de los mismos, pudiendo llegar a afectar incluso la economía nacional. Por tal motivo, se hace necesario darle seguimiento a toda la información relacionada con el consumo, producción, extracción y comercialización de estos portadores, con el propósito de evitar el malgasto de estos recursos de gran valor para la sociedad.

A continuación, se explican brevemente las clasificaciones que pueden recibir los portadores energéticos de acuerdo a su origen, conjuntamente con algunos ejemplos de los cuales se recoge información en la ONE.

Portadores energéticos primarios o naturales: son aquellos que se obtienen de la naturaleza de forma directa, a través de la fotosíntesis o después de atravesar un proceso minero. [18] Algunos de ellos son:

- Petróleo
- Gas natural
- Leña
- Productos de la caña

Portadores energéticos secundarios o artificiales: son aquellos productos resultantes de las transformaciones o de un proceso de elaboración a partir de portadores energéticos naturales, y en determinados casos obtenidos a partir de otro portador ya elaborado. [18] Algunos ejemplos son:

- Electricidad
- Carbón vegetal

- Alcohol desnaturalizado
- Gas manufacturado
- Derivados del petróleo

Toda la información relacionada con los portadores energéticos utilizados en Cuba, es recogida en la ONE a través de dos modelos. El primero es el modelo 5073 o Balance de consumo de portadores energéticos, donde se informa mensualmente sobre la comercialización y el consumo de portadores energéticos en cada empresa del país. El segundo modelo es el 0006 o Indicadores seleccionados, donde se recogen los datos concernientes a la producción y extracción de estos recursos, con una periodicidad mensual y anual.

El estudio de la información recogida en cada una de las empresas, es realizado por un grupo de especialistas que radican en la sede central de la ONE, con el propósito de obtener estadísticas de calidad que constituyan un mecanismo de control para los órganos administrativos del país.

2.3. Necesidades de información

En la presente investigación se propone dar solución a un grupo de problemas existentes en la ONE, relacionados con el manejo de la información referente al área de energía del SIEN. Por tal motivo, es de vital importancia identificar las necesidades de información que poseen los especialistas de dicha área, pues constituyen la base para un correcto diseño del mercado de datos.

Luego de varias entrevistas realizadas a los especialistas de la ONE, se decide clasificar la información en cinco grupos fundamentales: Consumo, Servicentro, Producción, Extracción e Indicadores anuales, donde los dos primeros se encuentran asociados al modelo 5073 y los tres últimos al modelo 0006. Se identificaron además, un conjunto de requisitos de información asociados a cada uno de los grupos definidos, los cuales se enumeran a continuación:

Consumo

RI1. Obtener el inventario inicial físico para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI2. Obtener las compras realizadas a CUPET para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI3. Obtener otras entradas para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI4. Obtener el consumo directo para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI5. Obtener el consumo indirecto para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI6. Obtener otras salidas para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI7. Obtener el inventario final físico para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI8. Obtener las compras al cargar para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI9. Obtener el total recibido del que efectúa la carga para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI10. Obtener el consumo utilizando tarjetas magnéticas para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI11. Obtener la cantidad entregada para consumo a través de tarjetas magnéticas para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI12. Obtener el saldo final de las tarjetas magnéticas para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI13. Obtener la carga para el próximo mes para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI14. Obtener el consumo acumulado realmente para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI15. Obtener el consumo acumulado el año anterior al que se informa para los portadores que se consumen dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

Servicentro

RI16. Obtener el inventario inicial físico para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI17. Obtener las compras realizadas a CUPET para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI18. Obtener otras entradas para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI19. Obtener el consumo directo para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI20. Obtener el consumo indirecto para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI21. Obtener otras salidas para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI22. Obtener el inventario final físico para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI23. Obtener las compras al cargar para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI24. Obtener el total recibido del que efectúa la carga para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI25. Obtener el consumo utilizando tarjetas magnéticas para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI26. Obtener la cantidad entregada para consumo a través de tarjetas magnéticas para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI27. Obtener el saldo final de las tarjetas magnéticas para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI28. Obtener la carga para el próximo mes para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI29. Obtener el consumo acumulado realmente para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

RI30. Obtener el consumo acumulado el año anterior al que se informa para los portadores que se comercializan dado un tiempo por: DPA, NAE, CAE, ORG y entidad.

Extracción

RI31. Obtener las cifras oficiales del plan acumulado hasta el período que se informa, de los portadores que se extraen.

RI32. Obtener las cifras de la ejecución real del mes que se informa, de los portadores que se extraen.

RI33. Obtener las cifras de la ejecución real del año actual, acumuladas hasta el período que se informa, de los portadores que se extraen.

RI34. Obtener las cifras de la ejecución real del año anterior, acumuladas hasta el período que se informa, de los portadores que se extraen.

Producción

RI35. Obtener las cifras oficiales del plan acumulado hasta el período que se informa, de los portadores que se producen.

RI36. Obtener las cifras de la ejecución real del mes que se informa, de los portadores que se producen.

RI37. Obtener las cifras de la ejecución real del año actual, acumuladas hasta el período que se informa, de los portadores que se producen.

RI38. Obtener las cifras de la ejecución real del año anterior, acumuladas hasta el período que se informa, de los portadores que se producen.

Indicadores anuales

RI39. Obtener las cifras oficiales del plan acumulado hasta el período que se informa, de los indicadores anuales.

RI40. Obtener las cifras de la ejecución real del año actual, acumuladas hasta el período que se informa, de los indicadores anuales.

RI41. Obtener las cifras de la ejecución real del año anterior, acumuladas hasta el período que se informa, de los indicadores anuales.

2.4. Requisitos funcionales

Para lograr satisfacer las necesidades del cliente definidas con anterioridad, se hace necesario identificar las funcionalidades que debe poseer el sistema. Estas funcionalidades son conocidas como requisitos funcionales de la aplicación y se definen durante la etapa de análisis.

Seguidamente se enumeran los requisitos funcionales de la solución:

RF1. Autenticar usuario.

RF2. Extraer datos.

RF3. Realizar transformación y carga de los datos.

RF4. Insertar usuario

RF5. Eliminar usuario

RF6. Insertar rol

RF7. Eliminar rol

RF8. Insertar reporte

RF9. Modificar reporte

RF10. Eliminar reporte

RF11. Configurar elementos del cubo

RF12. Ordenar resultados

RF13. Mostrar padres

RF14. Mostrar propiedades

RF15. Suprimir filas

RF16. Invertir eje

RF17. Detallar miembros

RF18. Entrar en detalles

RF19. Mostrar datos de origen

RF20. Mostrar gráfico

RF21. Configurar gráfico

RF22. Editar consulta MDX

RF23. Exportar reporte con formato '.pdf'

RF24. Exportar reporte con formato '.xls'

RF25. Imprimir reporte

2.5. Requisitos no funcionales

Los requerimientos no funcionales, representan aquellas características del sistema que le reportan al cliente alguna utilidad, como el rendimiento, la robustez y la fiabilidad, dándole más confianza y seguridad en la aplicación. Luego de un análisis profundo, se definieron 27 requisitos no funcionales para el mercado de datos, los cuales están descritos en el artefacto Especificación de requisitos. (Ver expediente del proyecto).

2.6. Casos de uso del sistema

Durante la fase de análisis y diseño de un almacén de datos, se definen los casos de uso del sistema, que tienen como objetivo describir la relación entre la aplicación y los usuarios. En ellos, se agrupan todos los requisitos funcionales y de información que hayan sido identificados con anterioridad, proporcionando una mejor comprensión del sistema. En la siguiente figura se muestra el Diagrama de Casos de Uso del Sistema (DCUS) para la solución.

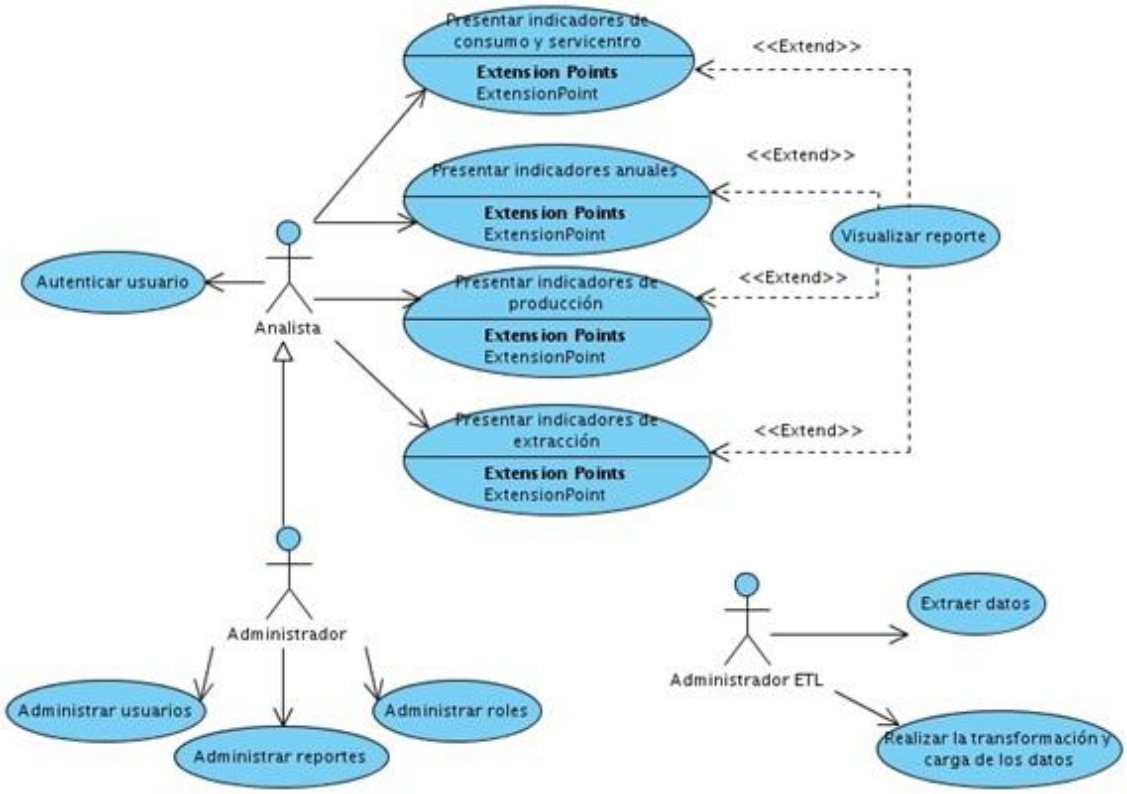


Figura 5: DCUS

Un caso de uso informativo, agrupa los requisitos de información que se definen teniendo en cuenta las necesidades del cliente. Por su parte, los casos de uso funcionales se encargan de agrupar las funcionalidades que debe tener el sistema. Seguidamente, se brinda una pequeña explicación de cada uno de ellos:

Casos de uso informativos

- **Presentar indicadores de consumo y servicentro:** comprende toda la información referente al consumo y comercialización de los portadores energéticos en cada empresa del país.
- **Presentar indicadores de extracción:** contempla los datos correspondientes a los portadores energéticos que son extraídos en el país.
- **Presentar indicadores de producción:** abarca toda la información relacionada con los portadores que se producen en el país.
- **Presentar indicadores anuales:** abarca toda la información relacionada con los portadores que analizan anualmente.

Casos de uso funcionales

- **Extraer datos:** se refiere al proceso durante el cual se extraen los datos de las fuentes de información.
- **Realizar la transformación y carga de los datos:** se refiere a los procesos de transformación y carga por el que pasan los datos, luego de ser extraídos de las fuentes.
- **Administrar roles:** abarca todos los requisitos relacionados con la administración de los roles.
- **Administrar reportes:** se refiere a la administración de todos los reportes relacionados con los portadores energéticos.
- **Administrar usuarios:** se refiere a la administración de los usuarios que tendrán acceso a la información contenida en el almacén.
- **Autenticar usuario:** consiste en el proceso de autenticación por el que deben pasar los usuarios, para controlar el acceso a los datos.
- **Visualizar reporte:** abarca todas las funcionalidades a través de las cuales un especialista puede modificar los reportes existentes de acuerdo a sus necesidades.

A continuación se detalla el funcionamiento del caso de uso Presentar indicadores de producción debido a su importancia en el mercado de datos:

Caso de Uso:	Presentar indicadores de producción
Tipo:	Información
Actores:	Analista
Resumen:	El caso de uso inicia cuando el especialista desea consultar la información

	relacionada con la producción mensual de portadores energéticos desde diferentes perspectivas de análisis. Luego de seleccionar el reporte deseado, el sistema muestra la información contenida en él y las opciones de los posibles cambios que le puede hacer al reporte. El caso de uso finaliza cuando el especialista termina de analizar la información relacionada con la producción mensual de portadores energéticos.
Precondiciones:	- La información relacionada con la producción mensual de portadores energéticos se cargó correctamente en el mercado de datos. - Todos los reportes relacionados con la producción mensual de portadores energéticos fueron creados.
Referencias	RI35, RI36, RI37, RI38. CU Visualizar reporte
Prioridad	Crítico

Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1. El especialista se autentica en el sistema.	2. Muestra la interfaz principal con las áreas de análisis existentes.
3. El especialista selecciona el área de análisis general A.A.G SIGOB.	4. Muestra las áreas de análisis que están contenidas dentro del A.A.G SIGOB.
5. El especialista selecciona el área de análisis A.A Control de energía.	6. Muestra los libros de trabajo que están contenidos dentro del A.A Control de energía.
7. El especialista selecciona el libro de trabajo L.T Producción.	8. Muestra los reportes contendidos dentro del L.T Producción.
9. El especialista selecciona el reporte deseado.	10. Muestra la información contenida en el reporte seleccionado y brinda la posibilidad al especialista de hacerle cambios al reporte para su análisis. Ir al Caso de Uso: Visualizar reporte. Finaliza el caso de uso.

Opciones de reportes de Indicadores de producción		
Entradas	**Posibles resultados**	
	Salidas	**Periodicidad**
Variables de entrada relacionadas con el	Variables de salida disponibles en el	Rango de tiempo en

<table>
<tr>
<td>caso de uso Presentar indicadores de producción.

➢ Producción portadores
➢ DPA
➢ Entidad
➢ Organismos
➢ NAE
➢ CAE
➢ Temporal mes</td>
<td>caso de uso Presentar indicadores de producción.

➢ Plan
➢ Real del mes
➢ Hasta este mes
➢ Real año anterior
➢ Estimado
➢ Dinámica
➢ Promedio mensual
➢ Diferencia dinámica
➢ Diferencia real-Estimado
➢ Valor absoluto dinámica
➢ Variación</td>
<td>que se solicitan las variables de salidas.

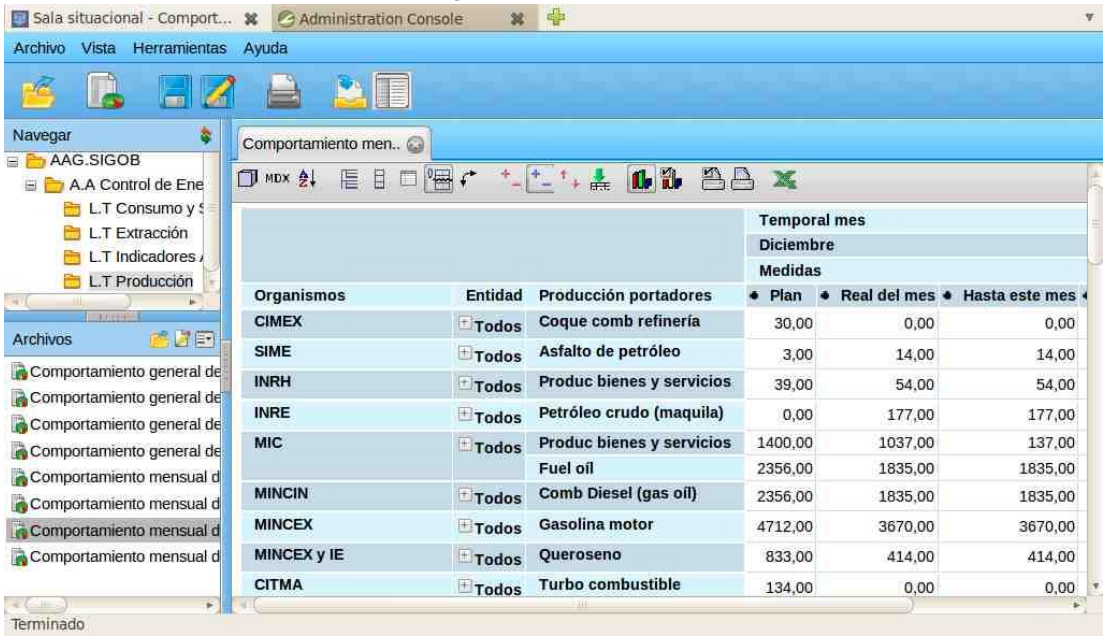
✓ Mensual</td>
</tr>
<tr>
<td colspan="3">Prototipo de interfaz de usuario
 
</td>
</tr>
<tr>
<td>Poscondiciones</td>
<td colspan="2">Los reportes correspondientes al libro de trabajo L.T Producción han sido consultados por el especialista.</td>
</tr>
</table>

Tabla 3: Caso de uso de información Presentar indicadores de producción

2.7. Modelo de datos dimensional

El modelado dimensional, es el nuevo nombre por el que se conoce a una vieja técnica para lograr que las bases de datos fueran más simples y entendibles, pero ha llegado a ser ampliamente aceptada como la técnica dominante para la presentación de los almacenes de datos. Además, ha surgido como la única arquitectura coherente para la construcción de estos sistemas, demostrando ser la estrategia más efectiva en cuanto al costo. El modelado dimensional, también resulta beneficioso en cuanto a diseño, pues posibilita una mejor comprensión de la aplicación por parte de los usuarios, un mayor rendimiento en las consultas y flexibilidad ante los cambios. [2]

En el modelo de datos, se definen las dimensiones y hechos que serán las futuras tablas de la base de datos de la solución. Debido a esto, se hace necesario conocer cada uno de sus componentes, los cuales se explican a continuación con el propósito de facilitar su entendimiento:

Hechos

En un modelo dimensional, una tabla de hechos representa una transacción o un evento del negocio y en ella se almacenan un conjunto de medidas o atributos, que permiten cuantificar o medir el rendimiento en los diferentes procesos del mismo. Generalmente, estas tablas poseen su propia llave primaria que se forma por la unión de las llaves pertenecientes a las dimensiones que se relacionan con ella, por lo que también se conoce como llave compuesta. [2]

Dimensiones

Por su parte, las tablas de dimensiones tienden a ser poco profundas en cuanto a la cantidad de filas, pero suelen ser bastante amplias en cuanto al número de columnas o atributos. Estos últimos, son la fuente principal de las restricciones de las consultas y los agrupamientos, por lo que en una consulta o una solicitud de reporte, son los diferentes aspectos por los que se pueden analizar las medidas de los hechos. Además, cada dimensión debe tener una llave primaria que sirva como base para la integridad referencial con cualquier tabla de hechos con la que se relacione. [2]

A continuación, se muestra el modelo de datos diseñado para la solución el cual está dividido en cuatro partes, para representar mejor las relaciones entre los hechos y las dimensiones:

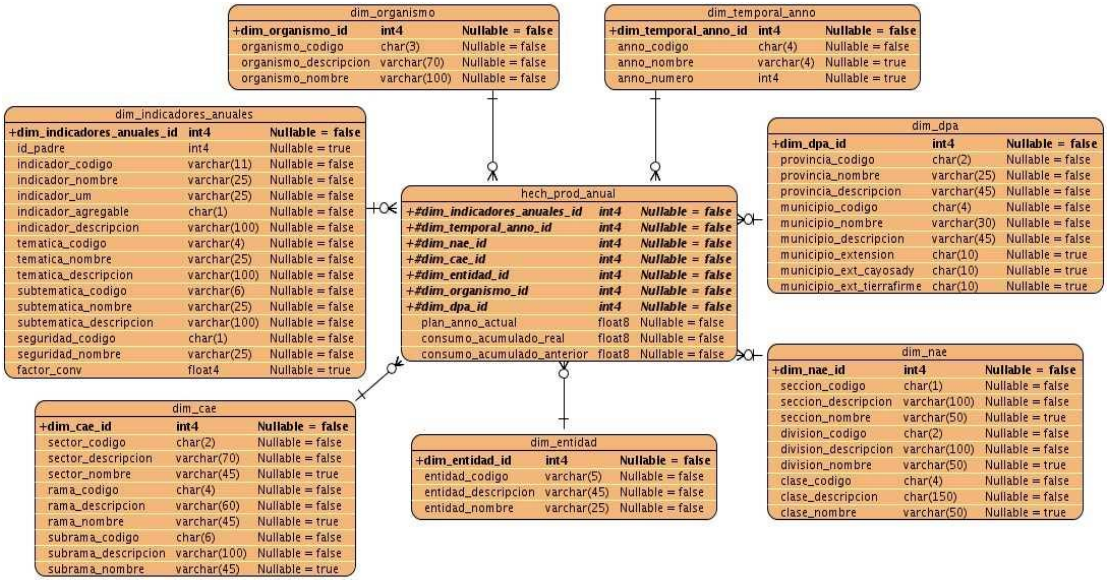


Figura 6: Modelo de datos dimensional. Hecho producción anual

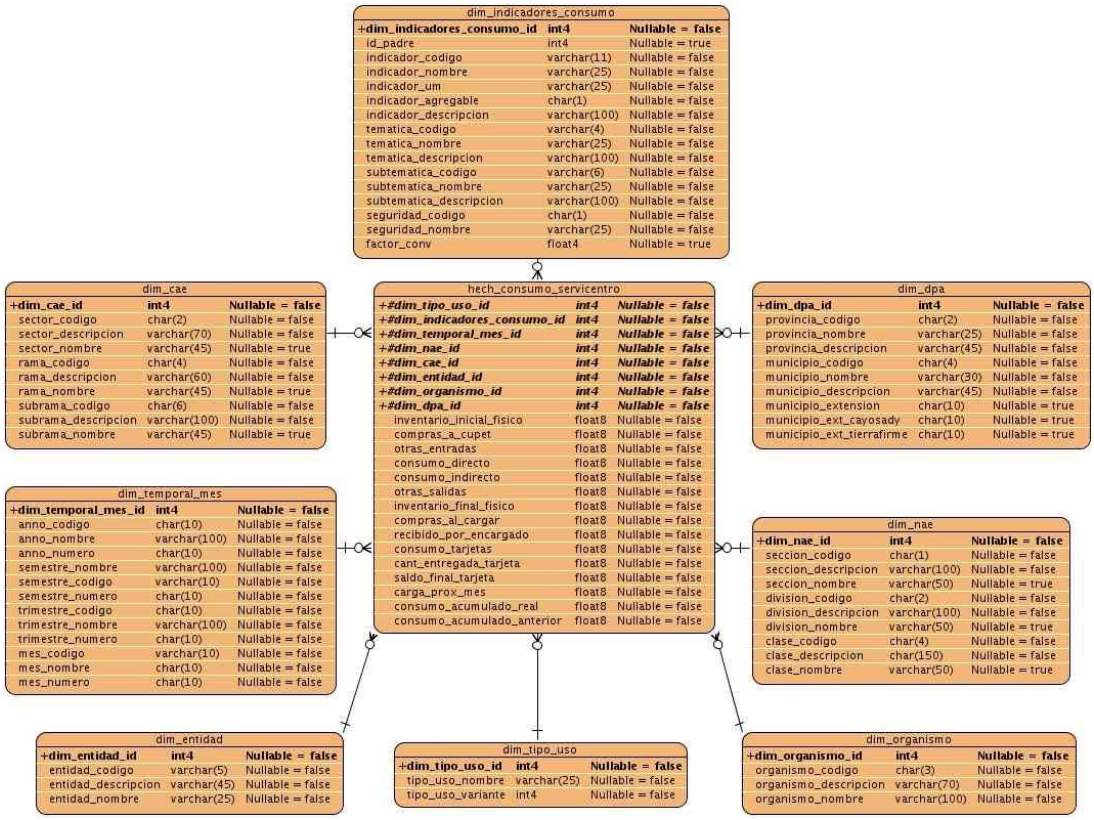


Figura 7: Modelo de datos dimensional. Hecho consumo y servicentro

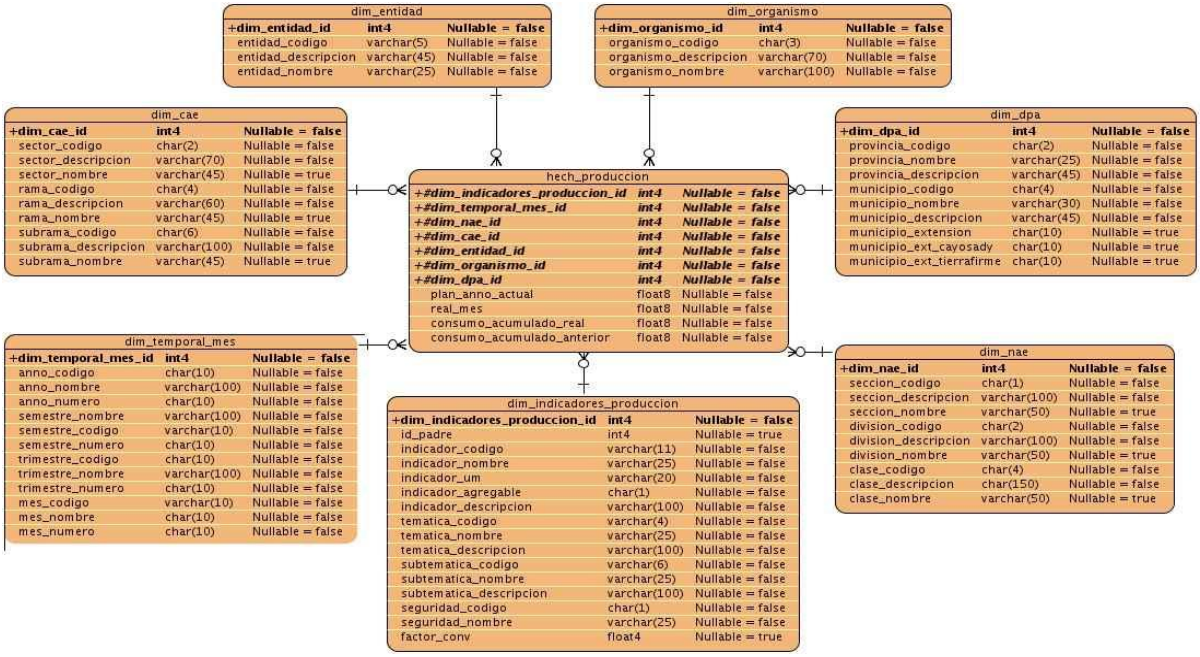


Figura 8: Modelo de datos dimensional. Hecho producción

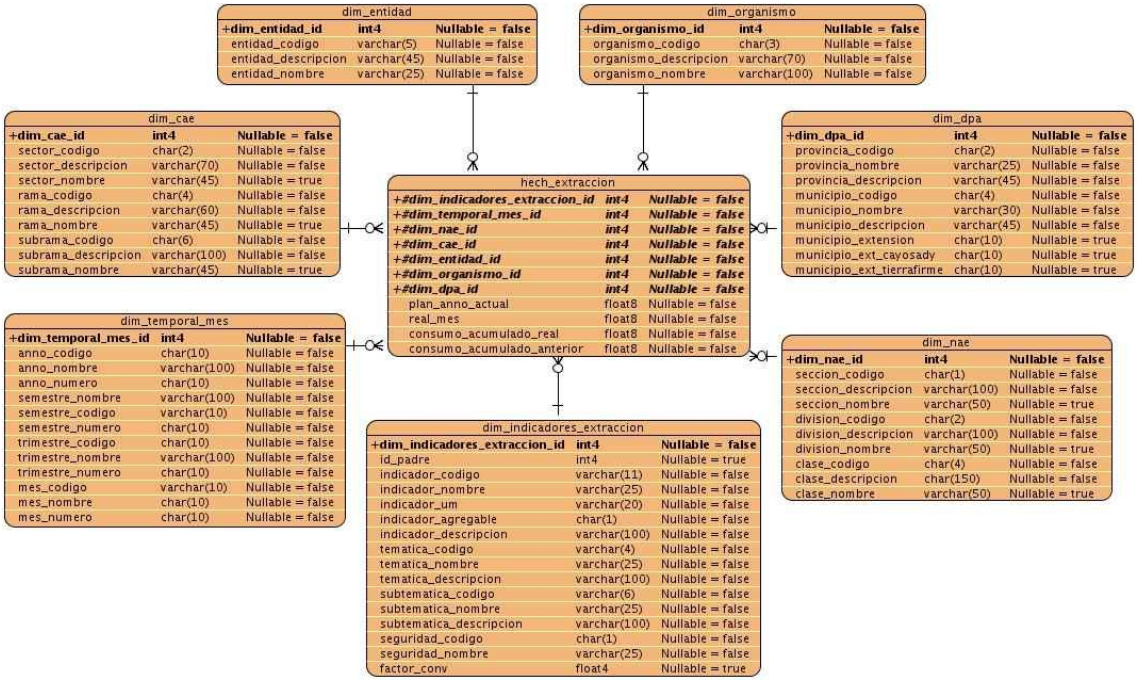


Figura 9: Modelo de datos dimensional. Hecho extracción

2.7.1. Identificación de dimensiones y hechos

Luego de un estudio realizado sobre toda la información que se recoge en la ONE relacionada con los portadores energéticos, fueron identificados los siguientes hechos y dimensiones:

Lista de hechos:

- **hech_consumo_servicentro:** agrupa los indicadores recogidos en el modelo 5073, relacionados con el consumo y comercialización de portadores energéticos.
- **hech_extraccion:** agrupa los indicadores recogidos en el modelo 0006 relacionados con la extracción de portadores.
- **hech_produccion:** agrupa los indicadores recogidos en el modelo 0006 relacionados con la producción de portadores.
- **hecho_prod_anual:** agrupa los indicadores recogidos en el modelo 0006 relacionados con los portadores que se analizan anualmente.

Lista de dimensiones:

- **dim_cae:** agrupa información relacionada con el Clasificador de Actividades Económicas y presenta tres niveles: sector, rama y subrama.
- **dim_dpa:** agrupa información relacionada con la División Política Administrativa de Cuba, compuesta por 14 provincias y 168 municipios.
- **dim_entidad:** agrupa información relacionada con todas las empresas cubanas.
- **dim_nae:** agrupa información relacionada con el Nomenclador de Actividades Económicas y presenta tres niveles: sección, división y clase.
- **dim_organismo:** agrupa información relacionada con todos los organismos cubanos.
- **dim_indicadores_consumo:** agrupa información relacionada con todos los portadores energéticos que se consumen en el país.
- **dim_tipo_uso:** está relacionada con la variante del modelo 5073 que puede ser consumo o servicentro.
- **dim_temporal_mes:** agrupa información relacionada con la periodicidad con que se recoge la información de los modelos.
- **dim_indicadores_extraccion:** agrupa información relacionada con los portadores energéticos que se extraen en el país.

- **dim_indicadores_produccion:** agrupa información relacionada con los portadores energéticos que se producen en el país.
- **dim_indicadores_anuales:** agrupa información relacionada con los portadores energéticos analizan anualmente.
- **dim_temporal_anno:** agrupa información relacionada con la periodicidad con que se recoge la información de los modelos.

Desarrollo de la matriz BUS

La matriz BUS o dimensional, representa la relación que existe entre las tablas de hechos y las tablas de dimensiones con las que se relaciona. Seguidamente, se muestra la matriz correspondiente al mercado de datos Control de energía, donde se reflejan las relaciones entre los hechos y las dimensiones explicadas anteriormente.

MATRIZ BUS				
DIMENSIONES	HECHOS			
	consumo_servicentro	extraccion	produccion	prod_anual
cae	X	X	X	X
dpa	X	X	X	X
entidad	X	X	X	X
nae	X	X	X	X
organismo	X	X	X	X
indicadores_consumo	X			
tipo_uso	X			
temporal_mes	X	X	X	
indicadores_extraccion		X		
indicadores_produccion			X	
indicadores_anuales				X
temporal_anno				X

Figura 10: Matriz BUS

2.8. Reglas del negocio

Las reglas del negocio, son políticas que deben cumplirse o condiciones que deben satisfacerse, las cuales regulan algún proceso del negocio. [19] Estas normas o restricciones, son definidas por los especialistas de

energía de la ONE de acuerdo a sus necesidades de información. A continuación, se muestran las reglas del negocio identificadas para el mercado de datos:

Reglas de portadores energéticos

Los siguientes indicadores tendrán un código de fila propio y serán calculados sumando los valores de los portadores que posean el código de fila correspondiente.

Indicador	Código de fila del indicador	Código de fila del portador correspondiente
Gasolina Motor Total	089	040, 041, 042
Combustible Diesel Total	189	120, 140
Aceites Lubricantes	289	220, 235, 260, 275
Aceites y Grasas Lubricantes	299	200, 220, 235, 260, 275
GLP total	389	360, 370
Petróleo Crudo Total	489	400, 430, 450, 470
Petróleo Combustible Total	589	500, 520, 540, 560
Mezclas Total	679	630, 640, 660
Fuel + Crudo + Mezclas	699	489, 589, 679
Nafta Total	749	700, 715, 730
Asfalto de Petróleo Total	799	750, 765, 780

Tabla 4: Reglas del negocio

Reglas generales de los modelos

Para ambos modelos, se crearán las siguientes variables en correspondencia con las necesidades de los especialistas del área de energía de la ONE:

- **Estimado:** se obtiene dividiendo el acumulado del mes anterior entre el número del mes anterior, más el acumulado del mes anterior.

- **Diferencia real estimado:** se obtiene restando el estimado menos el acumulado real del mes actual.
- **Dinámica o crecimiento:** se obtiene dividiendo el acumulado real del mes entre el acumulado del año anterior hasta el mes que se informa y multiplicando el resultado por 100.
- **Diferencia dinámica:** se obtiene restando el valor de la dinámica del mes anterior al que se informa menos el valor de la dinámica del mes actual.
- **Promedio mensual:** se obtiene dividiendo el acumulado del mes actual entre el número del mes.
- **Valor absoluto de la dinámica:** se obtiene restando el valor de la dinámica menos 100.
- **Variación:** se obtiene dividiendo la diferencia entre el acumulado real del mes y el acumulado del año anterior.
- **Variante:** es utilizada para identificar las variantes del modelo 5073, que puede tomar el valor uno si se reporta el consumo y dos si es servicentro.
- **DPA:** se representa con cuatro dígitos, los dos primeros dígitos pertenecen a la provincia y los dos siguientes identifican el municipio dentro de la provincia.
- **NAE:** presenta tres niveles donde el primero es la sección, identificada por una letra del alfabeto, seguido de cuatro dígitos de los cuales los dos primeros representan el segundo nivel que es la división y los dos últimos se corresponden con el tercer nivel que es la clase.
- **CAE:** se representa con seis dígitos donde los dos primeros dígitos pertenecen al sector, los dos siguientes identifican la rama dentro de ese sector y los dos últimos pertenecen a la subrama.

2.9. Seguridad del mercado de datos

Todo sistema de información, debe contar con un mecanismo de protección contra aquellas acciones que puedan afectar la integridad de los datos. Por tal motivo, la seguridad estará garantizada a través de la autenticación de los usuarios, a los cuales se les asigna un rol con un grupo de permisos para interactuar con la aplicación.

Seguidamente se muestran los roles y permisos definidos para la solución:

Roles	Permisos
Administrador	Tiene acceso total para gestionar los permisos, roles y usuarios, pero solo tiene acceso de lectura a las Áreas de Análisis (AA) del mercado de datos.
Analista	Tiene acceso de solo lectura al AA Energía. También,

	puede visualizar todos los reportes de esta área.
Administrador ETL	Tiene acceso total a todas las Áreas de Análisis (AA) del mercado de datos.

Tabla 5: Roles y permisos para el acceso a la información

Además de los roles y permisos definidos, se cuenta con la seguridad que proporciona la Plataforma Pentaho BI, la cual posee un mecanismo de seguridad compuesto por cuatro áreas fundamentales que son: [20]

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

2.10. Estrategias de recuperación y respaldo

Para garantizar la persistencia de la información, se realizará un respaldo total de los datos del Almacén de Datos con una frecuencia mensual y anual. Toda esta información, se almacenará en el área de la dirección de informática en un banco de datos especial. También, cada seis meses antes de comenzar los períodos vacacionales de julio y diciembre, se realizará una salva general de los datos en tres versiones (abuelo, padre e hijo) y se guardará en un DVD. Esta información, se almacenará en el edificio correspondiente a la oficina de estadísticas de La Habana y será responsabilidad del grupo de administración de redes de la ONE.

2.11. Conclusiones

En el capítulo, se abordaron los principales elementos relacionados con los artefactos generados durante la etapa de análisis y diseño del mercado de datos. En primer lugar, se realizó un estudio de la información que es analizada en el área de energía de la ONE, luego de lo cual se identificaron los requisitos de información, funcionales y no funcionales del sistema. Además, se definió el DCUS y el modelo de datos dimensional donde se reflejan las dimensiones y hechos identificados, así como también se definieron las reglas del negocio conjuntamente con el cliente. Finalmente, se explicó cómo se manejará la seguridad de los datos en la aplicación y las estrategias a seguir para la recuperación de los datos, en caso de ocurrir algún incidente que atente contra la integridad de los mismos.

Capítulo 3: Implementación del mercado de datos

3.1. Introducción

El capítulo está dirigido a la implementación de los diferentes aspectos relacionados con los procesos de integración de datos, con el propósito de brindar una mayor comprensión de las estrategias y procedimientos utilizados. Además, se abordan elementos relacionados con la implementación de la capa de inteligencia de negocio, incluyendo la creación de las estructuras necesarias para la navegación y el análisis de los datos.

3.2. Implementación del modelo de datos

Durante la etapa de análisis y diseño del mercado de datos, fueron definidos los hechos y las dimensiones que se convertirían en las tablas de la base de datos. Para el mercado de datos Control de energía se definieron dos esquemas principales, el primero para almacenar las dimensiones comunes a todos los departamentos de la ONE y un segundo esquema para las dimensiones y hechos específicos del área de energía, además de cuatro tablas auxiliares necesarias para la visualización de la información.

A continuación se pueden apreciar dichos esquemas con sus tablas correspondientes:

Esquemas	Tablas que lo componen
dimensiones	dim_cae dim_nae dim_dpa dim_organismo dim_entidad dim_temporal_mes dim_temporal_anno
mart_energía	dim_indicadores_anuales dim_indicadores_consumo dim_indicadores_extraccion dim_indicadores_produccion dim_tipo_uso hech_consumo_servicentro hech_produccion hech_extraccion hech_prod_anual closure_indicador_anual closure_indicador_consumo closure_indicador_produccion closure_indicador_extraccion

Tabla 6: Esquemas y tablas de la base de datos

3.3. Implementación del subsistema de integración de datos

El proceso de integración de datos se divide en tres etapas o fases fundamentales estrechamente relacionadas entre sí, las cuales son explicadas brevemente a continuación [21]:

Extracción: el primer paso a seguir es la exitosa extracción de los datos de los sistemas fuentes, donde cada uno posee un conjunto de características propias que necesitan ser manejadas correctamente. Una extracción efectiva es la clave para el éxito de un DWH, por lo que es necesario prestarle especial atención y formular una adecuada estrategia para realizar esta etapa.

Transformación: durante la fase de transformación, se aplican una serie de reglas del negocio o funciones sobre los datos extraídos, para convertirlos en datos con la calidad requerida por los usuarios finales. La eliminación de datos duplicados, la corrección de errores ortográficos, la aclaración de datos ambiguos y la estandarización de códigos, son algunas de las tareas que se realizan en esta fase del proceso de integración, que contribuyen a la limpieza y homogeneización de la información.

Carga: la fase de carga es el momento en el cual los datos obtenidos de la fase anterior, son cargados en el área de presentación, donde podrán ser consultados por los usuarios finales.

Con frecuencia, se puede observar la presencia de valores nulos en las fuentes de datos, que pudieran confundir al usuario mientras se encuentre analizando la información. Esto ocurre debido a que dichos valores no son los suficientemente descriptivos, por lo que no resultan de gran utilidad cuando se desea tomar una decisión que los involucre. Por tal motivo, se hace necesario reemplazarlos por valores más representativos para el usuario, teniendo cuidado de no alterar su significado para el negocio.

La estrategia a seguir para el tratamiento de este tipo de valores en el mercado de datos, se puede apreciar en la Figura 11 correspondiente a la transformación de la dimensión dim_indicadores_produccion. En este caso, se decide reemplazar el valor nulo que venga en el campo indicador_agregable por el valor “s” que indica que el portador es agregable, ya que todos los indicadores del mercado de datos lo son. Por su parte, el resto de los campos son sustituidos por el valor “Desconocido”, ya que no son de gran importancia para identificar a un portador.

#	Field	Replace by value
1	indicador_agregable	s
2	indicador_descripcion	Desconocido
3	indicador_nombre	Desconocido
4	indicador_um	Desconocido
5	subtematica_descripcion	Desconocido
6	subtematica_nombre	Desconocido
7	tematica_descripcion	Desconocido
8	tematica_nombre	Desconocido

Figura 11: Reemplazar valores nulos

A continuación, se describe de forma general la estrategia de integración definida para realizar los procesos de ETL, correspondientes a los hechos y dimensiones del mercado de datos, véase Figura 12.

Primeramente, se extraen los datos de las fuentes conformadas por ficheros Excel para las dimensiones y por ficheros DBF para los hechos. Una vez efectuada la extracción de los datos, se realizan las validaciones necesarias teniendo en cuenta las reglas del negocio identificadas y en caso de encontrarse valores incorrectos, el flujo se desvía hacia un fichero donde es almacenado con la correspondiente descripción del error. Si se comprueba que los datos poseen la calidad requerida, se les realiza un conjunto de transformaciones y se procede a su inserción en el mercado de datos.

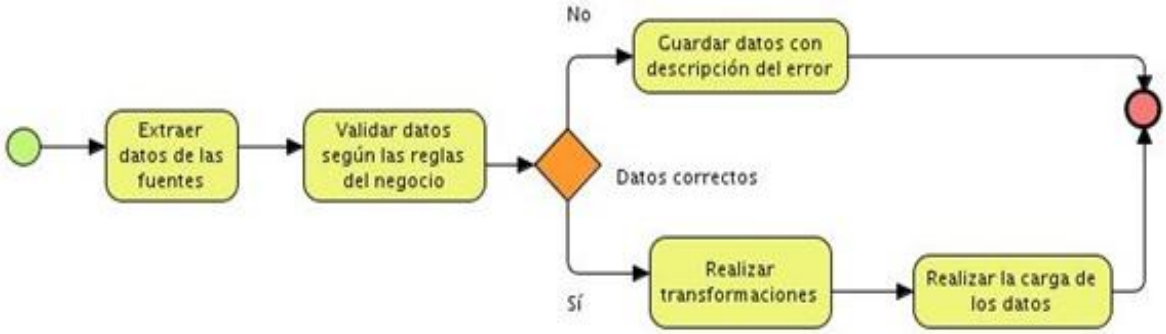


Figura 12: Diseño general de las transformaciones

Transformaciones para las dimensiones

Para realizar la extracción de los datos correspondientes a cada una de las dimensiones relacionadas con los portadores energéticos, se accede a un fichero con formato Excel que contiene los campos necesarios y se procede a realizar las transformaciones pertinentes. Como se puede apreciar en la Figura 13 correspondiente

a la transformación de la dimensión dim_indicadores_produccion, en un paso posterior se procede al tratamiento de los campos obtenidos de la fuente teniendo en cuenta las particularidades de cada uno.

Un ejemplo claro lo constituye el mapeo de valores, a través del cual se consigue un formato único y homogéneo para un campo determinado, así como la sustitución de los valores nulos por una etiqueta preestablecida, evitando de esta forma que existan inconsistencias en el mercado de datos. Además, durante la validación el flujo es desviado hacia un paso de error, donde se tratan los datos cuyos errores no pudieron ser corregidos durante las transformaciones anteriores. En caso de que la validación sea un éxito, se procede a la carga inicial de los datos en la tabla de dimensión.

En el caso de las dimensiones relacionadas con los portadores, existe un identificador para el portador padre, quien representa un nivel superior en la jerarquía. En una carga inicial, se insertan todos los datos relacionados con los portadores, a excepción del identificador del portador padre, debido a que para su carga resulta necesaria la existencia de todos los datos en la dimensión. En un paso posterior, se filtran los portadores que tienen un padre definido en la fuente y se realiza una búsqueda en la tabla de dimensión, para verificar que dicho código pertenezca a un portador válido ya insertado. Finalmente, se obtiene el identificador correspondiente al portador en cuestión y se realiza la segunda y última carga, donde se actualiza el identificador del portador padre para las entradas que así lo requieran.

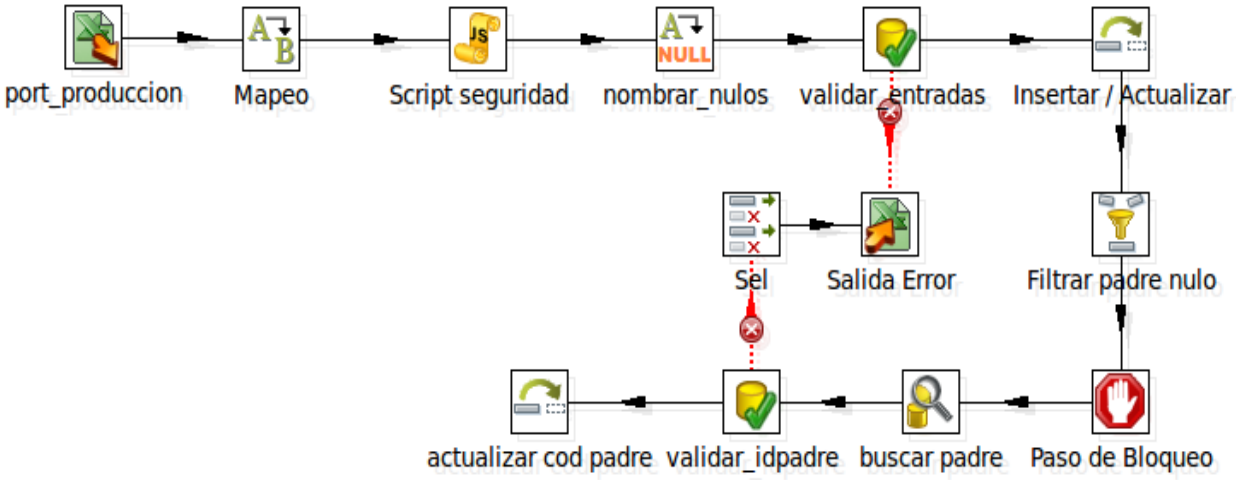


Figura 13: Transformación de la dimensión dim_indicadores_produccion

Transformaciones para los hechos

Para realizar la explicación de los procesos de integración de los hechos, se toma como ejemplo la transformación correspondiente al hecho hech_consumo_servicentro. En la Figura 14, se muestra como se

realiza la extracción y validación de los datos del fichero con formato DBF relacionado con el modelo 5073, donde se recoge información de los portadores que se consumen y comercializan en el país.

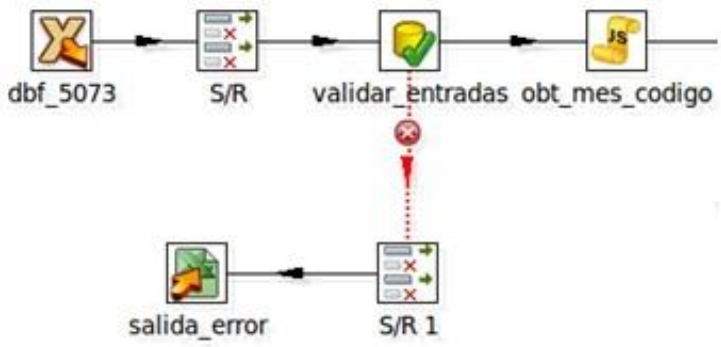


Figura 14: Extracción y validación de los datos

Seguidamente, se realiza una búsqueda de los identificadores de cada dimensión que se relacione con el hecho. Es de especial atención la dimensión dim_indicadores_produccion, debido a que el código que se encuentra en la fuente no es el mismo almacenado en la base de datos, por lo cual se utiliza un fichero con formato Excel que contiene un mapeo de ambos códigos, permitiendo finalmente acceder al identificador de dicha dimensión. Una vez obtenidos todos los identificadores necesarios, se chequea que sus valores no sean nulos para evitar la inserción de entradas inconsistentes, que violen las restricciones de integridad definidas en el hecho. Finalmente, después de comprobar que los datos estén correctos, se procede a realizar la carga de los datos en la tabla (Ver Figura 15).

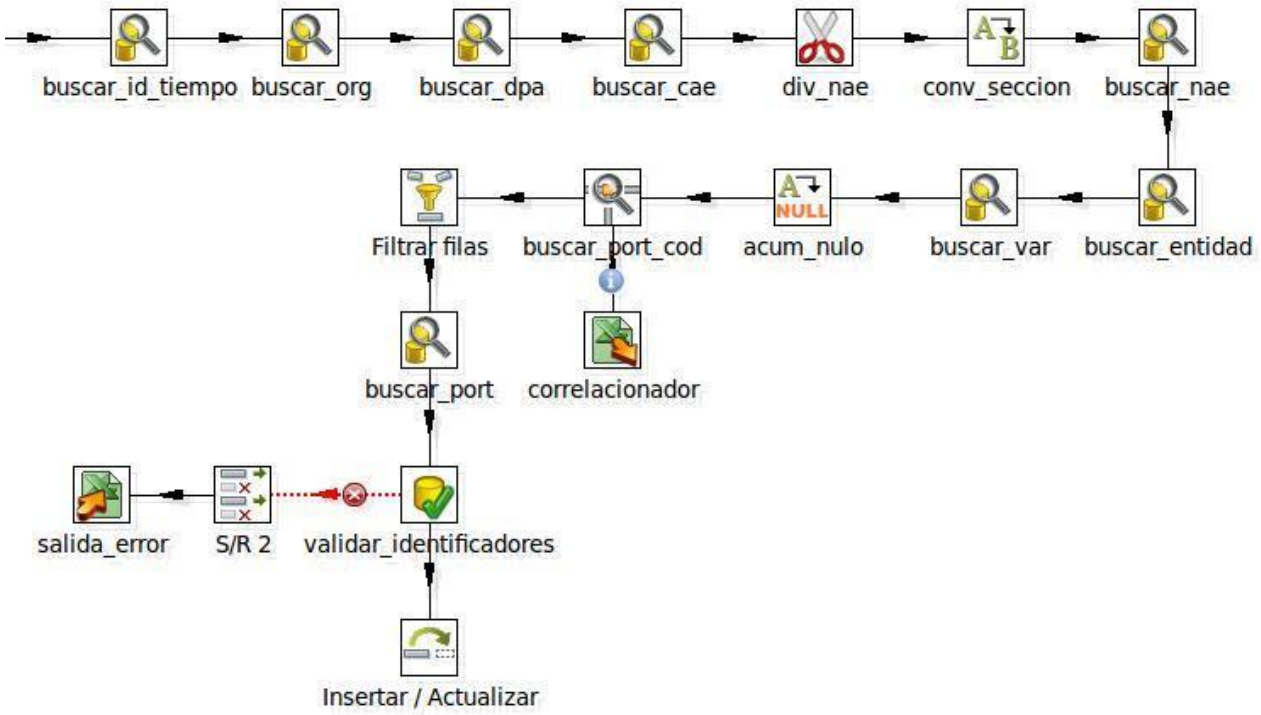


Figura 15: Búsqueda de dimensiones y carga de los datos

Flujo principal

Con el propósito de definir una secuencia lógica para la ejecución de las transformaciones, se crea un trabajo principal (ver Figura 17), donde primeramente se ejecuta un trabajo mediante el cual se cargan las dimensiones y las tablas auxiliares (ver Figura 18).

Figura 17: Trabajo general

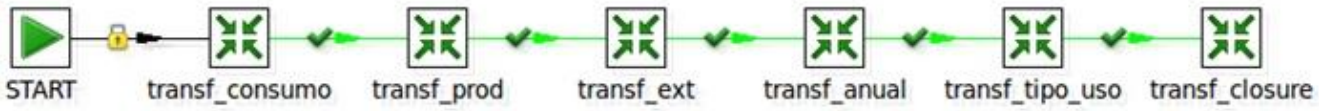


Figura 18: Trabajo para la carga de las tablas de dimensiones y tablas auxiliares

En un segundo paso, se realiza una transformación para obtener todos los ficheros fuentes que van a ser cargados en el mercado de datos (ver Figura 19) y finalmente, se ejecuta un trabajo para cargar las tablas de hechos de acuerdo a los ficheros que existan (ver Figura 20).

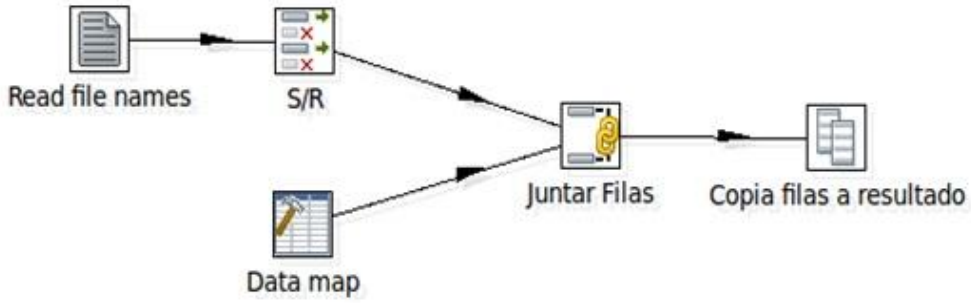


Figura 19: Trabajo para obtener los ficheros a cargar

Figura 20: Trabajo para cargar las tablas de hechos

Debido a la dependencia que poseen las tablas de hechos con las tablas de dimensiones, se hace necesario ejecutar primero las transformaciones correspondientes a las dimensiones y posteriormente las relacionadas con los hechos.

3.4. Implementación del subsistema de visualización de datos

En correspondencia con los requisitos de información identificados para el área de energía, se definieron 40 reportes agrupados en cuatro Libros de Trabajo (LT) ubicados dentro del Área de Análisis (AA) Control de energía. Dicha área se corresponde con una sección del almacén de datos del proyecto SIGOB, mientras que los LT representan las diferentes categorías a las que pueden pertenecer los reportes, ver Figura 21.

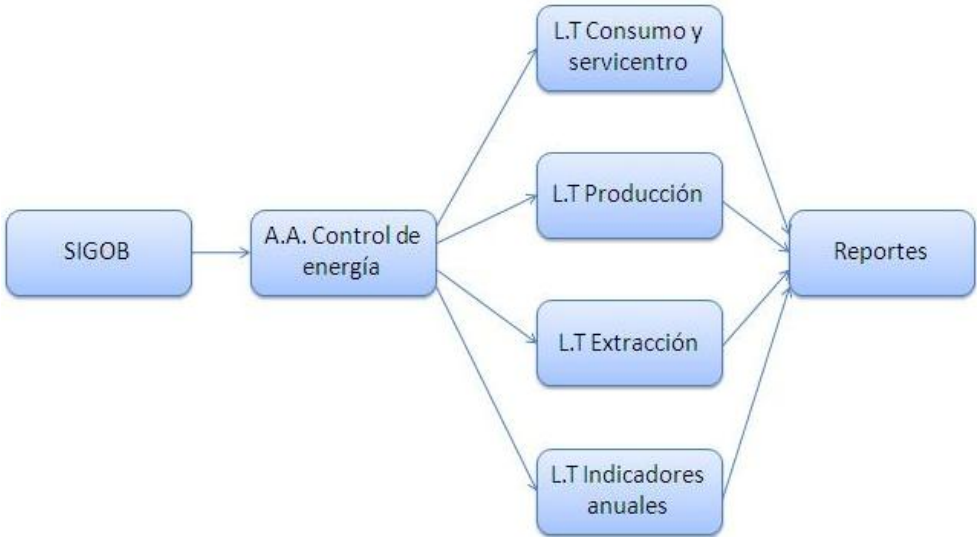


Figura 21: Mapa de navegación

3.4.1 Cubos OLAP

El total de hechos definidos para un mercado de datos, indica la cantidad de cubos OLAP que deben ser creados para la visualización de la información. En correspondencia con los hechos definidos para el área de energía, se implementaron los siguientes cubos OLAP:

Figura 22: Cubo OLAP correspondiente al hecho hech_prod_anual

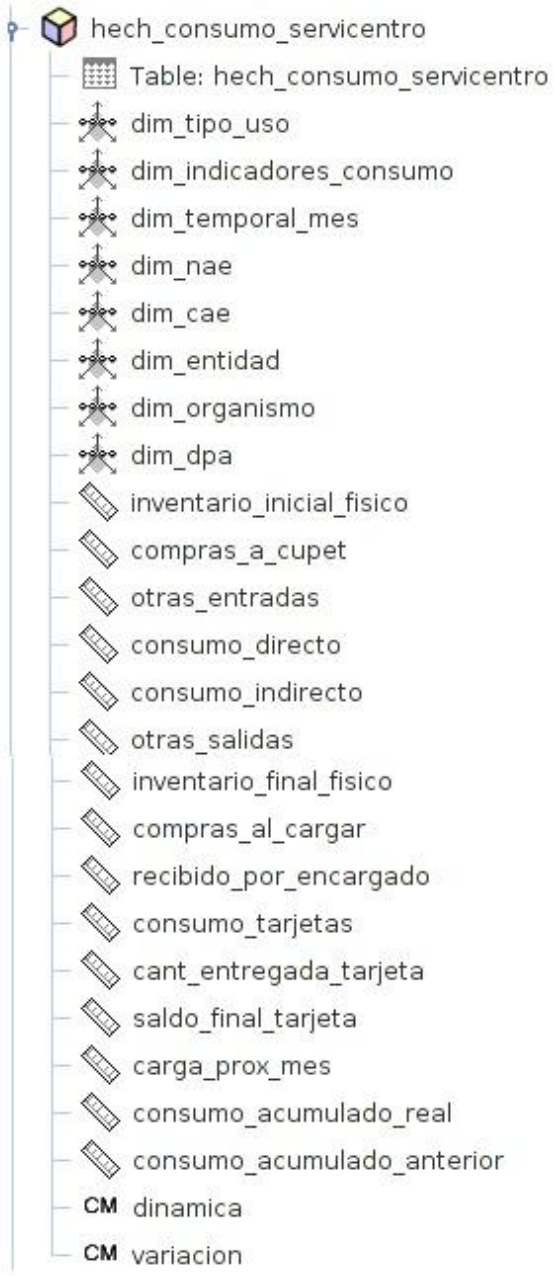


Figura 23: Cubo OLAP correspondiente al hecho hech_consumo_servicentro

Figura 24: Cubo OLAP correspondiente al hecho hech_extracción

Figura 25: Cubo OLAP correspondiente al hecho hech_produccion

3.5. Conclusiones

En el capítulo se abordaron los elementos más importantes relacionados con la implementación del subsistema de ETL, donde se explicaron las transformaciones realizadas para la carga de los datos en las tablas de dimensiones y hechos, así como el orden en que deben ser ejecutadas. Además, se expusieron aspectos de interés relacionados con la organización y visualización de la información, como son la creación de los cubos OLAP y la estructura definida para la navegación.

Capítulo 4: Validación del mercado de datos

4.1. Introducción

En el capítulo se exponen las pruebas realizadas al mercado de datos, así como los resultados obtenidos en cada una de ellas luego de su aplicación. Dichas pruebas son realizadas para garantizar el cumplimiento de las exigencias del cliente y la calidad del producto.

4.2. Pruebas de software

Debido a que resulta imposible que el hombre trabaje de forma perfecta, el proceso de desarrollo de un software debe ir acompañado de una actividad que garantice la calidad del resultado final. El término “prueba del software”, es un concepto bastante amplio que a menudo es conocido como verificación y validación. La verificación, se refiere al conjunto de actividades que aseguran que el software implementa correctamente una función específica, mientras que la validación se refiere a un conjunto diferente de actividades que aseguran que el software construido se ajusta a los requisitos del cliente. [22] Por esta razón, se hace necesario aplicar diferentes pruebas luego de culminada la etapa de implementación, con el propósito de verificar no solo que el producto cumpla con los requisitos del cliente, sino también para eliminar los posibles defectos que pudiera tener.

A continuación, se muestran algunas de las pruebas que pueden ser utilizadas para la validación de un software: [23]

Pruebas unitarias: esta prueba centra el proceso de verificación en la menor unidad del diseño del software: el componente de software o módulo.

Pruebas de integración: es una técnica sistemática cuyo objetivo consiste en probar el sistema como un todo, para detectar errores asociados con la interacción entre los diferentes módulos que lo componen.

Pruebas de validación: proporciona una seguridad final que el software satisface todos los requisitos funcionales, de comportamiento y de rendimiento.

Pruebas de regresión: consiste en volver a ejecutar un subconjunto de pruebas que se han llevado a cabo anteriormente, para asegurarse que los cambios que se hayan realizado no introduzcan un comportamiento no deseado o errores adicionales.

Pruebas de aceptación: se realizan para probar que el sistema cumpla con los requerimientos y expectativas del cliente.

Para el desarrollo de cualquier producto de software se realizan diferentes actividades desde que surge la idea inicial hasta la obtención del producto final. Para establecer un orden de ejecución de estas actividades de utilizó el modelo V, el cual es empleado por el centro DATEC para garantizar la calidad del producto final. En la Figura 26, se aprecia una representación gráfica del ciclo de vida del software propuesta en el modelo V, donde a la izquierda se muestran las diferentes etapas de desarrollo, mientras que a la derecha se muestran las pruebas correspondientes a cada una de ellas.

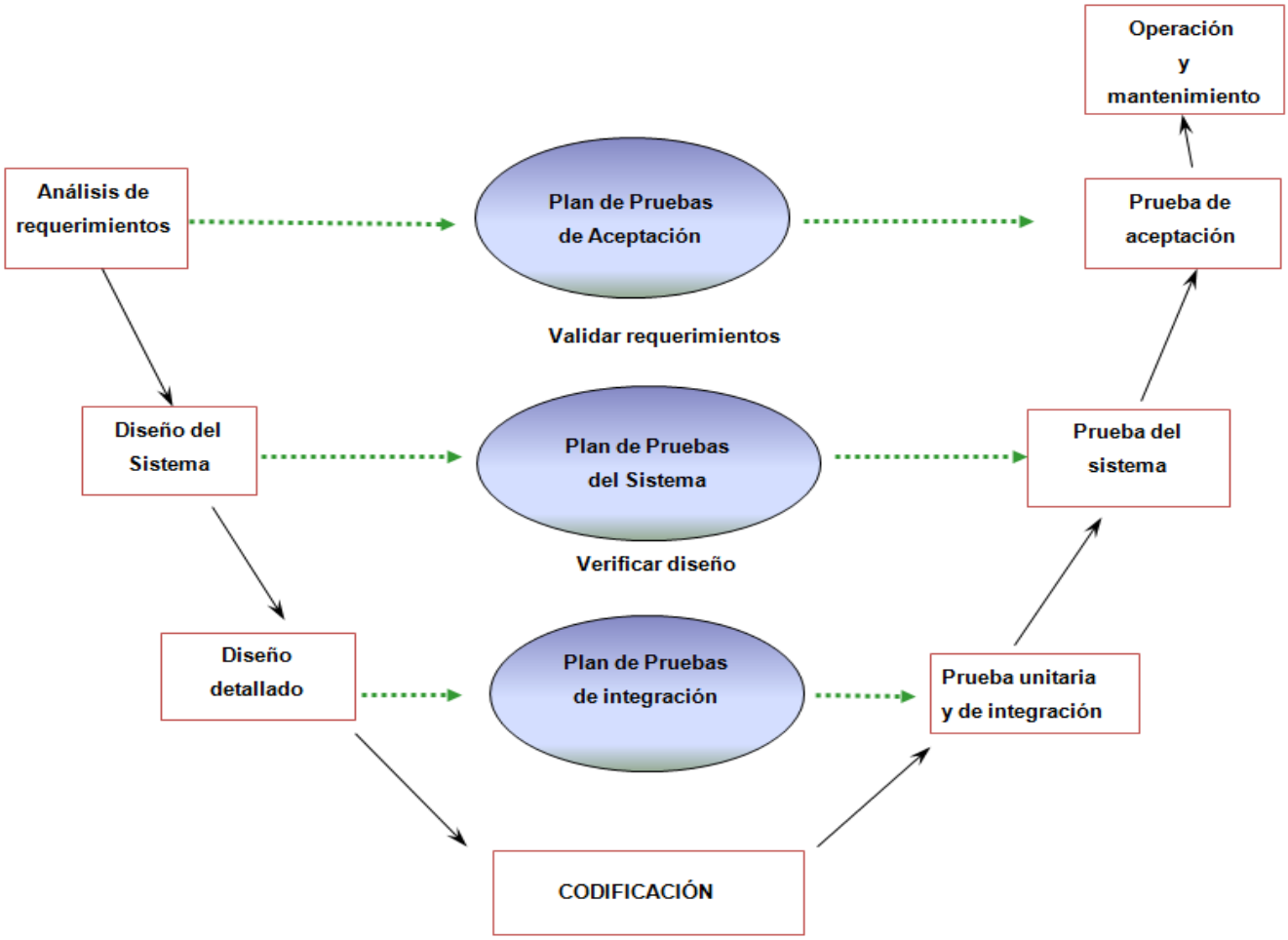


Figura 26: Modelo V

4.3. Herramientas para la aplicación de las pruebas

Para la validación del mercado de datos, se diseñaron cuatro casos de prueba, con el propósito de verificar los requisitos de información, agrupados en cuatro casos de uso de información que fueron definidos

previamente durante la etapa de análisis. Es importante destacar, que cada caso de uso puede estar asociado a más de un caso de prueba, en dependencia de la complejidad que posea.

En la siguiente tabla, se muestra un fragmento del caso de prueba correspondiente al caso de uso Presentar indicadores de producción, donde se aprecia el primer reporte correspondiente al LT Producción con una breve descripción, además de los perfiles de análisis correspondientes, los indicadores que se miden en el reporte, la respuesta que debe dar el sistema y el flujo central de ejecución del caso de prueba.

Escenario	Descripción	Perfiles de análisis	Indicadores a medir	Respuesta del sistema	Flujo central
EC 1.1: Comportamiento general de la industria en la producción por dpa, portador y entidad	Permite visualizar el reporte con las variables presentes en el mismo.	➢ Producción portadores ➢ DPA ➢ Entidad ➢ Temporal mes	➢ Hasta este mes ➢ Real año anterior ➢ Estimado ➢ Dinámica ➢ Promedio mensual ➢ Diferencia dinámica ➢ Diferencia real estimado ➢ Valor absoluto dinámica ➢ Variación	El sistema muestra todas las variables disponibles para los análisis, ubicados en las filas y las columnas que pueden ser visualizadas para cada reporte.	Se autentifica. Se entra al sistema. Se despliega hacia la derecha el componente ubicado en el lateral izquierdo que contiene el navegador.

Tabla 7: Caso de prueba correspondiente a los indicadores de producción

También, con el propósito de validar los artefactos desarrollados a partir de un grupo de indicadores, se utilizó una lista de chequeo general, previamente definida por el centro. Para una mayor organización, la lista de chequeo se divide en tres secciones que son: estructura del documento, indicadores definidos en el desarrollo y semántica del documento, que agrupan un total de 14 indicadores de los cuales ocho tienen un peso crítico.

4.4. Resultados de las pruebas

Seguidamente se detallará de forma breve cada una de las pruebas aplicadas de acuerdo a la propuesta del modelo V, así como los resultados obtenidos en cada una de ellas:

Pruebas unitarias y de integración

Una vez culminada la implementación se realizaron pruebas unitarias a los flujos de integración de datos y a los diferentes componentes relacionados con la capa de visualización, siendo detectadas cinco no conformidades que fueron resueltas rápidamente. Además, se realizaron pruebas internas por parte de los profesores del departamento Almacenes de datos, donde fueron identificadas 11 no conformidades

relacionadas con la estructura de los casos de prueba, y los reportes candidatos, las cuales fueron totalmente resueltas.

Pruebas de sistema

También, se realizaron pruebas por parte del grupo de especialistas de calidad del centro DATEC, donde fue identificada una no conformidad relacionada con una funcionalidad de la aplicación que no se ejecutaba correctamente, la cual fue igualmente resuelta. Del mismo modo se ejecutaron pruebas por parte del grupo de calidad de la UCI (CALISOFT), donde fueron detectadas 13 no conformidades relacionadas con la numeración de los reportes y los tipos de datos definidos para las variables en los casos de prueba, las cuales también fueron resueltas en el tiempo establecido para ello.

Pruebas ejecutadas aplicando listas de chequeo

De igual forma, luego de aplicada la lista de chequeo general a los principales artefactos generados durante la investigación, se detectaron tres no conformidades que también fueron resueltas en un período corto de tiempo. En la siguiente imagen, se puede apreciar el comportamiento de los indicadores medidos en la lista de chequeo, así como las no conformidades encontradas luego su aplicación.

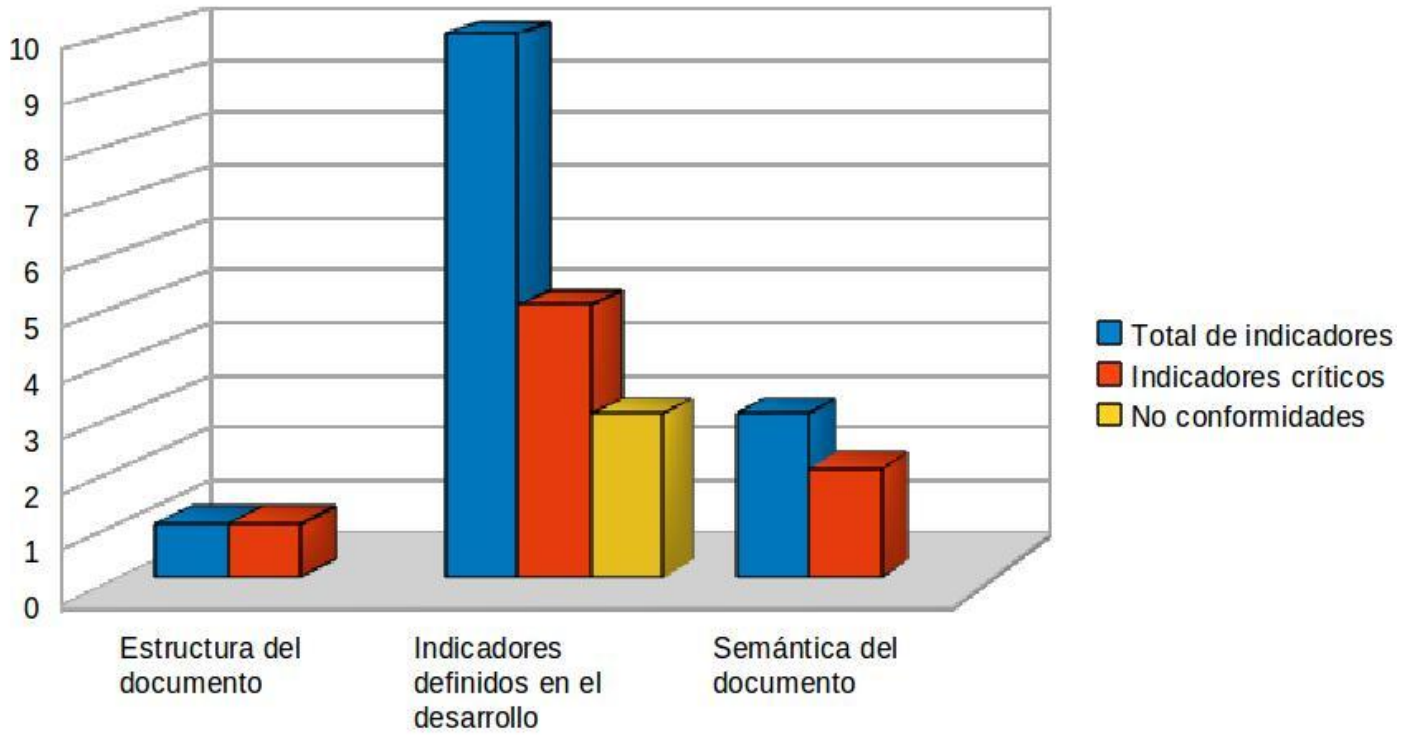


Figura 27: Comportamiento de los indicadores medidos en la lista de chequeo

Pruebas de aceptación

Una vez finalizado el desarrollo del mercado de datos, el cliente también puede probar el sistema para verificar no solo el cumplimiento de los requerimientos definidos, sino también el funcionamiento de la aplicación desde el punto de vista de un usuario final. Estas pruebas, demostrarán si el producto satisface realmente las necesidades de los especialistas de la ONE, ya que el personal que las ejecuta forma parte del equipo de trabajo que utilizará el mercado de datos en dicho centro.

Las pruebas de aceptación para el mercado de datos fueron realizadas por la especialista Elena Leonila Fernández García, quien es la representante de la ONE en la UCI. Los resultados arrojados luego de ejecutar dichas pruebas resultaron satisfactorios, ya que se confirmó que el mercado satisface todas las necesidades de información identificadas con anterioridad. Dichos resultados, han sido oficializados a través de una carta de aceptación firmada por la especialista, donde acepta que el producto está listo para ser utilizado.

4.5. Conclusiones

En el capítulo, se explicó brevemente la estrategia seguida para la validación del mercado de datos tomando como base la propuesta del modelo V. También, se detallaron las diferentes herramientas utilizadas para la validación del mercado de datos y se abordaron las diferentes pruebas realizadas para verificar el cumplimiento de los objetivos definidos, conjuntamente con los resultados obtenidos luego de ser aplicadas.

Conclusiones

Finalmente se puede concluir que:

El análisis realizado satisfizo las necesidades planteadas por los clientes finales del área de energía de la ONE.

El diseño modelado e implementado estuvo acorde con el análisis realizado en esta área y cumple con lo requerido para la posterior integración de los datos.

El proceso de extracción, transformación y carga (ETL) se implementó satisfactoriamente al insertar en el mercado de datos para el área Control de energía, los datos requeridos para los posteriores análisis de información.

La implementación de los reportes definidos responde a las necesidades de información del usuario final y permiten una eficiente y eficaz presentación de la información requerida, para la toma de decisiones.

Las pruebas realizadas permitieron validar el mercado de datos obteniendo resultados satisfactorios.

Recomendaciones

Se recomienda implementar una estrategia para realizar la carga de los ficheros de error obtenidos como resultado de las transformaciones, luego de ser analizados y corregidos por los especialistas del área de energía de la ONE.

Gestionar metadatos que registren información acerca de cada uno de los procesos de integración, permitiendo de esta forma facilitar al equipo de ETL, depurar los errores que puedan ocurrir de una forma más acertada, así como la generación de informes detallados sobre el estado de los datos.

Referencias bibliográficas

[1] Inmon, W. H. Building the Data Warehouse. s.l.: Wiley Publishing. ISBN: 0-471-08130-2.

[2] Kimball, Ralph y Ross, Margy. The Data Warehouse Toolkit. s.l.: Wiley Computer Publishing. ISBN: 0-471-20024-7.

[3] Ponniah, Paulraj. Data Warehousing Fundamentals. EUA: Wiley Publishing Inc, 2001. ISBN: 0-471-22162-7.

[4] Bouman, Roland y van Dongen, Jos. Pentaho Solutions, Business Intelligence and Data Warehousing with Pentaho and MySQL. s.l: Wiley Publishing. ISBN: 978-0-470-48432-6.

[5] Caramazana, Alberto. Tecnologías y Metodologías para la Construcción de Sistemas de Gestión del Conocimiento. Madrid: s.n., 2002.

[6] Smith, Armstrong. Darlene and Michel Oracle Discoverer 10g Handbook. San Francisco, California: The McGraw-Hill Companies, 2006.

[7] Inteligencia de negocio. [En línea] 2007. [Citado el: 16 de noviembre de 2010.] http://www.ibermatica.com/ibermatica/businessintelligence2.

[8] NEXTEL Engineering. La Inteligencia de Negocios. [En línea] [Citado el: 20 de noviembre de 2010.] http://www.nexteleng.es/microsoft/bi.asp.

[9] Zepeda Sánchez, Leonardo. Metodología para el Diseño Conceptual de Almacenes de Datos. Valencia: Universidad Politécnica de Valencia, 2008.

[10] Visual Paradigm. [En línea] 2010. [Citado el: 18 de noviembre de 2010.] http://www.visual-paradigm.com.

[11] Sistemas gestores de bases de datos. [En línea] 2007. [Citado el: 28 de noviembre de 2010.] http://www.desarrolloweb.com/articulos/sistemas-gestores-bases-datos.html.

[12] González, Carlos D. Curso Base de Datos PostgreSQL, SQL avanzado y PHP. [En línea] [Citado el: 29 de Noviembre de 2010.] http://www.usabilidadweb.com.ar/postgre.php.

[13] DataCleaner. http://datacleaner.eobjects.org (último acceso: 30 de Noviembre de 2010)

[14] Roldan, María Carina. Pentaho 3.2 data integration: beginner's guide. s.l.: Packt Publishing. ISBN: 9781847199546.

[15] Pentaho open source bussines intelligence. [En línea] 2010. [Citado el: 3 de diciembre de 2010.] http://www.pentaho.com.

[16] Pentaho open source bussines intelligence. [En línea] 2010. [Citado el: 3 de diciembre de 2010.] http://mondrian.pentaho.com.

[17] Chopra, Vivek, Li, Sing y Genender, Jeff. Professional Apache Tomcat 6. Indianapolis: Wiley Publishing, 2007. ISBN: 9780471753612.

[18] Oficina Nacional de Estadísticas. http://www.one.cu/aec2009/esp/10_tabla_cuadro.htm (último acceso: 21 de enero de 2011)

[19] IBM Rational Software, RUP_Help. www.rational.com (último acceso: 15 de febrero de 2011).

[20] Pentaho.http://wiki.pentaho.com/display/ServerDoc1x/03.+Platform+Security+Implementation (último acceso: 19 de febrero de 2011).

[21] Kimball, Ralph y Caserta, Joe. The Data Warehouse ETL Toolkit Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. s.l.: Wiley Publishing. ISBN: 0-764-57923-1.

[22] Pressman, R.S. Ingeniería del software. Un enfoque práctico. s.l.: McGraw-Hill. ISBN: 84-481-3214-9.

[23] Scalone, Fernanda. “Estudio comparativo de los modelos y estándares de calidad del software”. Universidad tecnológica nacional, Facultad regional de Buenos Aires. 2006.

Bibliografía

1. Oficina Nacional de Estadísticas, ONE. [En línea] [Citado el: 11 de octubre de 2010.] http://www.one.cu.
2. Definición.de. [En línea] [Citado el: 13 de octubre de 2010.] http://definicion.de/estadistica/.
3. Dataprix. [En línea] [Citado el: 20 de octubre de 2010.] http://www.dataprix.com/ca/aggregator/sources/70.
4. Sinnexus. [En línea] [Citado el: 13 de noviembre de 2010.] http://www.sinnexus.com/.
5. ETL-Tools.Info. [En línea] [Citado el: 15 de noviembre de 2010.] http://etl-tools.info/es/bi/proceso_etl.htm.
6. Pentaho Community. [En línea] [Citado el: 18 de noviembre de 2010.] http://wiki.pentaho.com.
7. Gravitar. [En línea] [Citado el: 21 de noviembre de 2010.] http://www.gravitar.biz/index.php/herramientas-bi/.
8. Visual Paradigm. [En línea] [Citado el: 22 de noviembre de 2010.] http://www.visual-paradigm.com/support/documents/vpumluserguide/12/13/5963_aboutvisualp.html.
9. Dataprix. [En línea] [Citado el: 25 de noviembre de 2010.] http://www.dataprix.com/category/integracion-datos/perfilado-datos.
10. Pentaho, open source business intelligence. [En línea] [Citado el: 4 de diciembre de 2010.] http://mondrian.pentaho.com/documentation/workbench.php.
11. DataCleaner. [En línea] [Citado el: 30 de Noviembre de 2010.] http://datacleaner.eobjects.org.
12. Kimball, Ralph y Caserta, Joe. The Data Warehouse ETL Toolkit Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. s.l.: Wiley Publishing. ISBN: 0-764-57923-1.
13. Kimball, Ralph, Reeves, Laura, Ross, Margy y Thorntwhaite, Warren. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses. s.l.: Wiley Publishing. ISBN: 978-0-471-25547-5.
14. Imhoff, Claudia, Galemmo, Nicholas y Geiger, Jonathan. Mastering Data Warehouse Design, Relational and Dimensional Techniques. s.l.: Wiley Publishing. ISBN: 0-471-32421-3.
15. Caramazana, Alberto. Tecnologías y Metodologías para la Construcción de Sistemas de Gestión del Conocimiento. Madrid: s.n., 2002.
16. Pentaho open source bussines intelligence. [En línea] 2010. [Citado el: 3 de diciembre de 2010.] http://mondrian.pentaho.com.
17. Roldan, María Carina. Pentaho 3.2 data integration: beginner's guide. s.l.: Packt Publishing. ISBN: 9781847199546.

18. Visual Paradigm. [En línea] 2010. [Citado el: 18 de noviembre de 2010.] http://www.visual-paradigm.com.
19. Pentaho open source bussines intelligence. [En línea] 2010. [Citado el: 3 de diciembre de 2010.] http://www.pentaho.com.
20. Inmon, W. H. Building the Data Warehouse. s.l.: Wiley Publishing. ISBN: 0-471-08130-2
21. Ponniah, Paulraj. Data Warehousing Fundamentals. EUA: Wiley Publishing Inc, 2001. ISBN: 0-471-22162-7.
22. Bouman, Roland y van Dongen, Jos. Pentaho Solutions, Business Intelligence and Data Warehousing with Pentaho and MySQL. s.l: Wiley Publishing. ISBN: 978-0-470-48432-6.
23. Pressman, R.S. Ingeniería del software. Un enfoque práctico. s.l.: McGraw-Hill. ISBN: 84-481-3214-9.
24. Scalone, Fernanda. “Estudio comparativo de los modelos y estándares de calidad del software”. Universidad tecnológica nacional, Facultad regional de Buenos Aires. 2006

Glosario de términos

ONE: Oficina Nacional de Estadísticas.

TIC: Tecnologías de la Información y las Comunicaciones.

SIEN: Sistema de Información Estadístico Nacional.

SIGOB: Sistema de Información de Gobierno.

DATEC: Centro de Tecnologías de Gestión de Datos.

Data Warehouse: almacén de datos.

Data Mart: mercado de datos.

ETL: proceso de extracción, transformación y carga.

BI: Inteligencia del negocio.

Staging area: es un área de almacenamiento temporal donde se realizan un conjunto de procesos comúnmente conocidos como extracción, transformación y carga.

Herramientas CASE: conjunto de aplicaciones informáticas orientadas al incremento de la productividad en el desarrollo de software, las siglas CASE vienen dadas por su nombre en inglés Computer Aided Software Engineering que se conoce como Ingeniería de Software Asistida por Computadoras.

UML: lenguaje visual para especificar, construir y documentar un sistema de software. Sus siglas vienen dadas por su nombre en inglés Unified Modeling Language.

XML: estándar de información cuyas siglas vienen dadas por su nombre en inglés eXtensible Markup Language.

DB2: gestor de bases de datos relacional.

Oracle: herramienta cliente/servidor para la gestión de bases de datos.

SQL: lenguaje de consulta estructurado o SQL (por sus siglas en inglés Structured Query Language) es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones en ellas.

API: una API o Interfaz de Programación de Aplicaciones, es el conjunto de funciones y procedimientos (o métodos si se refiere a programación orientada a objetos) que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción.

JDBC: es el acrónimo de Java Database Connectivity, una API que permite la ejecución de operaciones sobre bases de datos desde el lenguaje de programación Java, independientemente del sistema operativo donde se ejecute o de la base de datos a la cual se accede utilizando el dialecto SQL del modelo de base de datos que se utilice.

Granularidad: representa el nivel de detalle al que se desea almacenar la información y se define en dependencia del negocio que se esté analizando.

Lista de chequeo: instrumento de medición y evaluación que consiste básicamente en un formulario de preguntas referentes al atributo de calidad que se está probando y de las características del documento en el caso de la documentación.

No conformidad: defecto, error o sugerencia que se le hace al equipo de desarrollo una vez encontrada alguna dificultad en lo que se está evaluando.