

Universidad de las Ciencias Informáticas

Facultad 3

CEIGE



Mercado de datos de Energías Renovables a partir de datos abiertos para la recuperación y análisis de información

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

Autores:

Manuel Alejandro Fernández Rodríguez

Camilo José Tamayo Morales

Tutores:

Ing. Olga Yarisbel Rojas Grass

MSc. Yusniel Hidalgo Delgado

Ciudad de La Habana, junio de 2017



El mundo camina hacia la era de electrónica... Todo indica que esta ciencia se constituirá en algo así como una medida del desarrollo; quien la domine será un país de vanguardia. Vamos a volcar nuestros esfuerzos en este sentido con audacia revolucionaria"

Ernesto Che Guevara

DECLARACIÓN DE AUTORÍA

Declaramos ser los autores de este trabajo y autorizamos a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

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

Manuel Alejandro Fernández Rodríguez
Autor

Camilo José Tamayo Morales
Autor

Ing. Olga Yarisbel Rojas Grass
Tutor

MsC. Yusniel Hidalgo Delgado
Tutor

Síntesis de los Tutores:

Ing. Olga Yarisbel Rojas Grass.

Ingeniera en Ciencias Informáticas, UCI 2009.

Email: varisbel@uci.cu

MSc. Yusniel Hidalgo Delgado.

Ingeniero en Ciencias Informáticas, UCI 2010.

Email: yhdelgado@uci.cu

Camilo:

Dedico esta tesis a mis padres que sin su apoyo y sus consejos mi paso por la UCI no hubiera ocurrido tan bien como lo fue, además por saber educarme a lo largo de mi corta vida. A mi hermano Ernesto por siempre sacarme de quicio y a mi pequeñito hermano Lionel por darme su cariño como más nadie sabe.

Alejandro:

Dedico esta tesis a mis padres por siempre brindarme su apoyo y comprensión en estos cinco años y a lo largo de mi vida. A mi familia por siempre estar pendiente de mí y ayudarme en todo momento. A todos los que hicieron posible que llegara hasta aquí.

Camilo:

Agradezco a mis padres Camilo y Virgen por darme la mejor educación del mundo y por quererme de la forma que lo hacen.

Agradezco a mis 2 hermanos del alma Ernesto y Lionel que, aunque a veces nos comportamos como 3 niños pequeños, nos queremos como nadie más en el mundo.

Agradezco a mi familia en general por haber confiado en mí y por toda la ayuda que me ha brindado.

Agradezco a mi compañero de tesis Manuel que sin su ayuda este trabajo no hubiera progresado.

Agradezco a mis amistades que durante todo este tiempo han compartido conmigo momentos buenos y no tan buenos, a las personas que de una forma u otra me han brindado su ayuda durante estos 5 años. En especial a: Ángel, Adrián, Yasiel, Liomar, Yanet y a todos los que vinieron conmigo de la Facultad Regional de Ciego de Ávila.

Agradezco a mi tutora Olga por tener la paciencia que ha tenido con nosotros, a su esposo Victor por calmarla cuando estaba estresada y a su hijo Joseito.

Alejandro:

Agradezco:

A mis padres por haberme formado como persona y haberme apoyado a llegar hasta aquí sin importar la distancia.

A mi mamá por aguantar todos estos años sola en la casa y siempre preocuparse por mí y tenerme paciencia.

A mi familia por apoyarme en los malos momentos y ayudarme a olvidar estos.

A mi otra familia de 19 y 2, Pomares, Sara, Liuba, Willy, Danae por siempre estar pendiente de mí, brindarme su consejo y su cariño.

A mis amigos del barrio por los buenos momentos, divertirme con sus locuras y preocuparse por mí.

A mis compañeros de aula, los que llegamos hasta aquí y a los que ya no están, por ayudarme en todo. En especial a: Yoandry, Omar, Jessica, Raiko y los que me brindaron su sincera amistad.

A ti rubia, por divertirme, apoyarme, bríndame tu cariño y hacerme pasar muy buenos momentos.

A la persona que me apoyó gran parte de mi carrera aunque no hasta el final, igual gracias.

A mi tutora por la paciencia y ayudarnos a salir adelante.

A mi compañero de tesis Camilo por su apoyo y ayuda para culminar este trabajo.

A todos los que de alguna forma u otra contribuyeron a llegar hasta aquí.

El Centro de Gestión de la Información y Desarrollo de la Energía (CUBAENERGIA) tiene como misión gestionar información sobre las energías y realizar acciones de divulgación que contribuyan a potenciar la cultura científica en el país. Con el objetivo de apoyar al centro en la recuperación y análisis de datos publicados en diferentes fuentes sobre energías renovables tanto a nivel nacional como internacional se propone como solución un Mercado de datos. Para dar cumplimiento al objetivo propuesto, fue centralizada la información que se encuentra almacenada en los documentos Excel, cuyo análisis de manera independiente se torna difícil y complejo. Dicho mercado de datos contribuye al análisis de los datos publicados en diferentes fuentes sobre energías renovables para el apoyo a la toma de decisiones. En el proceso de desarrollo se utilizó la metodología realizada por el Departamento de Almacenes de Datos que toma como base la metodología de Kimball y se ajusta a las condiciones y características de producción de la Universidad de la Ciencias Informáticas. Las herramientas y tecnologías empleadas en el desarrollo de la solución fueron Visual Paradigm, PostgreSQL, Pentaho Data Integration, Pentaho Schema Workbench y Pentaho BI Server. Como resultado se obtiene la estructura del modelo de datos dimensional que comprende: las tablas de dimensiones, hechos y las medidas necesarias para la visualización de los datos. Se definen e implementan los subsistemas de almacenamiento, integración y visualización. Una vez concluidos dichos procesos, se obtiene como resultado un mercado de datos poblado y una aplicación web para la visualización de diferentes reportes.

Palabras Claves:

Almacén de datos, integración, inteligencia de negocio, mercado de datos.

Introducción	1
Capítulo 1: Fundamentación teórica	5
1.1 Introducción	5
1.2 Almacén de Datos	5
1.3 Proceso de Extracción, Transformación y Carga (ETL)	5
1.4 Inteligencia de Negocios	6
1.5 Mercado de Datos	6
1.6 Procesamiento Analítico en Línea (OLAP)	7
1.6.1 ROLAP	8
1.6.2 MOLAP	8
1.6.3 HOLAP	9
1.6.4 Valoración general	9
1.7 Modelado dimensional	10
1.8 Metodología para el desarrollo del Mercado de datos	13
1.8.1 Metodología para el desarrollo de Almacén de Datos propuesta por el Centro de Tecnologías de Gestión de Datos (DATEC)	13
1.8.2 Descripción de las actividades en cada fase	14
1.9 Conclusiones del capítulo	15
Capítulo 2: Análisis y diseño	16
2.1 Introducción	16
2.2 Necesidades del Negocio	16
2.3 Requisitos	17
2.3.1 Requisitos de información	17
2.3.2 Requisitos no funcionales	18
2.4 Casos de uso de información	19
2.5 Arquitectura de un mercado de datos	23
2.6 Diseño de la solución	24
2.6.1 Diseño del subsistema de almacenamiento	24
2.6.2 Diseño del subsistema de integración	28
2.6.3 Diseño del subsistema de visualización	29
2.7 Política de respaldo y recuperación	30
2.8 Esquema de seguridad	30
2.8.1 Seguridad en la base de datos	31

2.8.2	Seguridad en la aplicación.....	31
2.9	Conclusiones del capítulo.....	31
Capítulo 3: Implementación y validación.....		32
3.1	Introducción.....	32
3.2	Implementación del subsistema de almacenamiento.....	32
3.2.1	Esquemas y tablas.....	32
3.2.2	Restricciones y secuencias.....	33
3.3	Implementación del subsistema de integración.....	34
3.3.1	Implementación de las transformaciones.....	34
3.4	Implementación del subsistema de visualización.....	35
3.4.1	Cubos de datos.....	35
3.5	Pruebas.....	38
3.5.1	Pruebas de integración.....	38
3.5.2	Matriz de trazabilidad.....	44
3.6	Conclusiones del capítulo.....	44
Conclusiones generales.....		45
Recomendaciones.....		46
Referencias Bibliográficas.....		47
Anexos.....		48
Glosario de términos.....		52

Tabla 1. Comparativa entre los sistemas de almacenamiento OLAP [20].	9
Tabla 2. Fuentes de energías renovables	16
Tabla 3. Estudio de las fuentes de datos.....	17
Tabla 4. CUI1 Obtener la información de energía renovable nacional	20
Tabla 5. CUI2 Obtener la información de energía renovable internacional	22
Tabla 6. Matriz dimensional	27
Tabla 7. Permisos de los usuarios en la base datos	31
Tabla 8. Permisos de los usuarios en la aplicación	31

Figura 1. Esquema estrella	11
Figura 2. Esquema copo de nieve	12
Figura 3. Esquema constelación de hechos	12
Figura 4. Diagrama de casos de uso de información	20
Figura 5. Arquitectura del mercado de datos	24
Figura 6. Dimensión tipos de energías renovables.....	25
Figura 7. Dimensión dispositivos generadores.....	25
Figura 8. Dimensión país.....	26
Figura 9. Dimensión año	26
Figura 10. Hecho producción de energías renovables.....	27
Figura 11. Modelo de datos del AD	28
Figura 12. Diseño de la transformación para la carga de los datos	29
Figura 13. Diseño de la transformación para la carga de los datos en los hechos	29
Figura 14. Cubo producción de energía renovable.....	30
Figura 15. Estructura de datos	33
Figura 16. Transformación para la carga de la dimensión dimension_tipos _energias_renovables	34
Figura 17. Transformación para la carga del hecho hech_produccion_energias_renovables.....	35
Figura 18. Interfaz de la vista de análisis de Energías renovables	36
Figura 19. Interfaz de la vista de análisis Energía renovable desplegada	36
Figura 20. Obtener el ahorro de petróleo por dispositivo generador y año.....	37
Figura 21. Resultado de la consulta hecha al MD.....	40
Figura 22. Archivo fuente de hidroeléctricas.....	40
Figura 23. Resultado de la consulta	42
Figura 24. Dispositivos instalados por provincias.....	43
Figura 25. Energía producida por dispositivos instalados.	43
Figura 26. Matriz de trazabilidad.....	44
Figura 27. Vista principal de la aplicación	48
Figura 28. Gráfico de análisis regional.....	48
Figura 29. Comportamiento internacional de la cantidad de energía producida anualmente.	49
Figura 30. Gráfica de análisis de cantidad de energía producida anualmente.....	49
Figura 31. Gráfica de análisis de ahorro en petróleo.....	50
Figura 32. Gráfica de análisis de ahorro en petróleo seleccionando una provincia	50
Figura 33. Gráfica de consumo físico de la provincia.....	51

Introducción

El desarrollo de fuentes de energías estables y seguras, constituye un elemento estratégico para el desarrollo económico de cualquier país [1]. Las energías renovables son energías limpias que contribuyen a cuidar el medio ambiente, frente a los efectos contaminantes y el agotamiento de los combustibles fósiles. Estas proceden de fuentes naturales que son inagotables como el sol, el aire, el agua, la biomasa entre otras [2]. Su impacto ambiental es nulo en la emisión de gases de efecto invernadero como el Dióxido de Carbono (CO₂). Se consideran la energía solar, la eólica, la geotérmica, la hidráulica como energías renovables y también pueden incluirse en este grupo la biomasa y la energía mareomotriz [3].

En la actualidad, varios gobiernos están publicando miles de conjuntos de datos para que personas y organizaciones los puedan utilizar y como consecuencia la cantidad de aplicaciones basadas sobre datos abiertos está incrementándose. La publicación de datos en formatos abiertos, abre la puerta a la posibilidad que los mismos sean reutilizados en nuevos proyectos, que puedan combinarse con otras fuentes de datos y generar nuevas aplicaciones desarrolladas por el gobierno, las organizaciones y las empresas.

A nivel mundial y en el sector que corresponde a las energías renovables, se encuentra la Agencia Internacional de Energías Renovables (IRENA¹), que es una organización intergubernamental que apoya a los países en su transición hacia un futuro energético sostenible y sirve como la plataforma principal para la cooperación internacional, un centro de excelencia y un repositorio de políticas, tecnología, recursos y conocimiento financiero sobre energías renovables. IRENA promueve la adopción generalizada y el uso sostenible de todas las formas de energía renovable, incluida la bioenergía, la energía geotérmica, hidroeléctrica, oceánica, solar y eólica, en la búsqueda del desarrollo sostenible, el acceso a la energía, la seguridad energética y el crecimiento económico y la prosperidad bajos en carbono. En esta agencia se encuentran vinculados 150 países [4], los informes estadísticos que ha publicado contienen datos desde el 2006 hasta el 2016.

Cuba se encuentra incluida dentro de los países que produce y genera este tipo de energía y por ende se muestran datos estadísticos del país en los informes generados por IRENA. La Oficina Nacional de Estadística e Información (ONEI), creada por el artículo 31 del Decreto Ley No. 281 del 2 de febrero de

¹ IRENA: International Renewable Energy Agency por sus siglas en inglés.

2011, como resultado de la organización del Sistema de Información del Gobierno presenta entre sus informes, estadísticas sobre energías renovables.

La Publicación Anual de la ONEI del Inventario Nacional de Fuentes de Energía Renovables tiene como objetivo mostrar datos estadísticos seriadados del uso de las Fuentes Renovables de Energía en Cuba. La fuente de información utilizada es el Modelo 5077 del SIEN y aporta datos que pueden ser utilizados por los especialistas e investigadores del área energética y medioambiental. Se insertan tablas y gráficos que reflejan la disponibilidad y consumos de las principales fuentes de energía, así como la instalación de los dispositivos capaces de generar energía alternativa y su uso en los territorios. Esta publicación es elaborada por los especialistas de la Dirección de Industrias de la Oficina Nacional de Estadísticas. Se debe mencionar que los datos que están publicados en el sitio pertenecen a los años del 2005 hasta el 2013.

En el país, el Centro de Gestión de la Información y Desarrollo de la Energía (CUBAENERGIA) tiene como misión gestionar información en energía, aplicaciones nucleares y tecnologías de avanzada; desarrollar productos y tecnologías para la sostenibilidad energética y realizar acciones de divulgación que contribuyan a potenciar la cultura científica en el país. Este centro pertenece a la Agencia de Energía Nuclear y Tecnologías de Avanzada (AENTA) y adscrito al Ministerio de Ciencia, Tecnología y Medio Ambiente (CITMA), es una entidad pública presupuestada de investigación - desarrollo y servicios científico - técnicos en materia de energía y medio ambiente [5].

En un encuentro realizado entre especialistas de esta institución y el Grupo de Web Semántica, que pertenece a la Facultad 3 de la Universidad de las Ciencias Informáticas, se explicó la necesidad de poder localizar e integrar datos e información procedente de organismos e instituciones que se dedican a producir datos e informes relacionados con las fuentes renovables de energía. Estos proceden de fuentes confiables y actualizadas y pueden ser utilizados para la creación de informes, para realizar vigilancia tecnológica sobre el tema de las energías renovables e incluso ayudar en la toma de decisiones.

Para combinar datos de varias fuentes de forma tal que parezca que provienen de una misma fuente, hay que tener en cuenta los siguientes problemas de integración que existen: cada fuente puede tener un modelo de datos distinto y las distintas fuentes pueden utilizar varias formas para representar el mismo concepto. Las razones antes descritas traen como consecuencia que se dificulte la búsqueda, el acceso y en definitiva no se utilice la información de la forma adecuada, ya que no se encuentra centralizada y puede no estar disponible en el momento que se necesite.

Problema a resolver:

¿Cómo contribuir al análisis de los datos publicados en diferentes fuentes sobre energías renovables para el apoyo a la toma de decisiones?

Se define como **objeto de estudio** el proceso de desarrollo de Almacenes de datos, enmarcado en el **campo de acción** el proceso de desarrollo de Mercados de datos sobre energías renovables.

Objetivo general: Desarrollar un mercado de datos que permita la recuperación y análisis sobre datos publicados en diferentes fuentes sobre energías renovables.

Objetivos específicos:

OE1. Definir el marco teórico de la investigación mediante el estudio y el análisis de los principales referentes teóricos acerca de los mercados de datos para el desarrollo de la solución.

OE2. Realizar el análisis y diseño del mercado de datos de energías renovables.

OE3. Realizar la implementación del mercado de datos de energías renovables.

OE4. Validar la propuesta de solución mediante la aplicación de técnicas y pruebas.

Como **resultado** se obtendrá un mercado de datos poblado, así como la capa de visualización de los datos que permitirá realizar análisis sobre las energías renovables.

La presente investigación está guiada por **métodos científicos** los cuales son:

Métodos teóricos:

Histórico-lógico: se utiliza para identificar las tendencias actuales en relación a los Almacenes de Datos y los Mercados de Datos, permitiendo conocer y estudiar los trabajos realizados con mayor relevancia hasta el momento, vinculados con el tema de investigación.

Analítico-sintético: se utiliza para la descomposición del problema de la investigación en elementos por separado para la comprensión de su funcionamiento, en este caso para distinguir, extraer y unificar los elementos que forman parte del proceso de construcción del mercado de datos.

El trabajo está compuesto por 3 capítulos:

Capítulo 1: Fundamentación teórica

En este capítulo se definen conceptos fundamentales sobre los almacenes de datos y los mercados de datos, con sus principales características, elementos que los componen y se hace referencia al desarrollo de los procesos de extracción, transformación y carga de los datos. Además, se explica y fundamenta la selección de la metodología a utilizar.

Capítulo 2: Análisis y diseño

En este capítulo se presentarán los principales elementos con respecto a las etapas de análisis y diseño de la solución. Se identificarán los requisitos de información y los requisitos no funcionales para dar respuesta a las necesidades del cliente. Se diseñarán los subsistemas de almacenamiento, integración y visualización. En cada una de las fases se incluirán y justificarán las herramientas y tecnologías que serán utilizadas.

Capítulo 3: Implementación y validación

En este capítulo se hace referencia a la implementación de la solución, abordando específicamente cómo se realiza la misma en el subsistema de almacenamiento, integración y visualización, teniendo en cuenta los requisitos y necesidades del negocio. Una vez terminados de implementar los subsistemas, se dará paso a la validación de la propuesta de solución mediante la aplicación de pruebas, para determinar que la solución desarrollada cuente con la calidad requerida.

Capítulo 1: Fundamentación teórica

1.1 Introducción

En el presente capítulo se describen los principales conceptos relacionados con los Almacenes de Datos (AD) y los Mercados de Datos (MD). Se realiza un estudio de los sistemas de almacenamientos de datos OLAP². Se analiza la metodología para el desarrollo de Almacenes de datos establecida en la Universidad.

1.2 Almacén de Datos

Un AD es un conjunto de datos integrados orientados a una materia que varían con el tiempo y que no son transitorios, los cuales soportan el proceso de toma de decisiones de una administración. [6]

Ralph Kimball define los AD como "una copia de los datos de transacciones estructuradas de manera específica para la consulta y análisis" [7]. También los define como "la unión de todos los mercados de datos de una entidad" [7], donde los MD son un repositorio de información, similar a un AD, pero orientado a un área o departamento específico de la organización.

Por tanto, los AD tienen como objetivo principal almacenar y proveer a una organización de información relevante y a tiempo. Su utilización ayuda a la integración y homogenización de los datos. Permiten el acceso a la información, pueden utilizarse como una fuente de conocimiento y contribuyen en la toma de decisiones.

1.3 Proceso de Extracción, Transformación y Carga (ETL)

El proceso de ETL es la base sobre la cual se alimenta el AD. Si se diseña adecuadamente, extrae los datos de los sistemas de origen de datos, aplica diferentes reglas para aumentar la calidad y consistencia de los mismos, consolida la información proveniente de distintos sistemas, y finalmente carga la información en el AD en un formato acorde para la utilización por parte de las herramientas de análisis [8]. A continuación se describen estos procesos [9]:

- Extracción: acción de obtener la información deseada a partir de los datos almacenados en fuentes externas.
- Transformación: cualquier operación realizada sobre los datos para que puedan ser cargados en el AD o se puedan migrar de este a otra base de datos.
- Carga: consiste en almacenar los datos en la base de datos final.

² OLAP: por sus siglas en inglés, On-Line Analytical Processing.

El flujo de trabajo que representa a los procesos de ETL está compuesto por varias funcionalidades que a continuación se mencionan [9]:

- Identificación de la información relevante a los recursos externos.
- Extracción de la información identificada.
- Transportación de la información al área de procesamiento.
- Transformación de la información proveniente de múltiples recursos a un formato común.
- Limpieza de los datos entrados a la base de datos.
- Propagación de los datos al AD.

1.4 Inteligencia de Negocios

Luego de realizar el proceso ETL, la información que se encuentra en el AD ya está preparada para ser utilizada por herramientas de análisis. Estas son conocidas como las aplicaciones de Inteligencia de Negocios (BI³), los informes y aplicaciones de análisis proporcionan información útil a los usuarios. Las aplicaciones de BI incluyen un amplio espectro de tipos de informes y herramientas de análisis, que van desde informes simples de formato fijo a sofisticadas aplicaciones analíticas. [7]

La BI se puede definir como un conjunto de técnicas y herramientas tanto de gestión empresarial como de aplicación tecnológica. Permiten a partir de la formulación estratégica y teniendo como objetivo dar soporte a los procesos de planeación y control en las organizaciones, la extracción e integración de los datos que son generados como resultado de la operación de las diferentes áreas funcionales en una organización, su posterior procesamiento y distribución en forma de información [10].

Vista como procesos, la inteligencia de negocios en su dimensión técnica pretende realizar la integración de los datos generados por y para la organización. Esta sirve como entrada a diferentes procesos de gestión en cada uno de los niveles organizacionales, y la distribución de la información generada para los usuarios interesados [11].

1.5 Mercado de Datos

El MD se centra en satisfacer las necesidades de un área específica de negocio en una organización, cuya información contribuye a la toma de decisiones.

Ventajas de utilizar MD [12]:

- Poco volumen de datos.

³ BI: Business Intelligence, por sus siglas en inglés.

- Mayor rapidez de consulta.
- Consultas MDX sencillas.
- Facilidad para conservar los datos a través del tiempo.
- Fácil acceso a los datos que se necesitan frecuentemente.
- Facilidad de creación.

Algunas características a tener en cuenta de los MD son las que se mencionan a continuación [13]:

- Según las necesidades de los usuarios el diseño del mercado de datos se realiza siguiendo una estructura consistente.
- La información histórica que posee es mínima.
- Contiene el grado de granularidad necesaria.
- Da costos adicionales en hardware, software y accesos de red.
- Debido a que hay grupos de usuarios que solo acceden a un subconjunto preciso de datos, se hace más fácil el acceso a las herramientas de consulta y divide los datos para controlar mejores accesos.

La diferencia que existe entre un MD y un AD es su alcance. El MD está hecho para cubrir las necesidades de un grupo de trabajo o de un determinado departamento dentro de la organización. Es el almacén natural para los datos departamentales. En cambio, el ámbito del almacén de datos es la organización en su conjunto. Es el almacén natural para los datos corporativos comunes. [14]

1.6 Procesamiento Analítico en Línea (OLAP)

Los sistemas OLAP son bases de datos orientadas al procesamiento analítico, donde el acceso a los datos es de solo lectura y la operación más común es la consulta, con pocas inserciones, actualizaciones o eliminaciones. Este sistema permite acceder a grandes volúmenes de datos, de los que se puede extraer información útil. OLAP es un tipo de procesamiento de datos que se caracteriza, entre otras cosas, por permitir el análisis multidimensional de datos [15].

Una aplicación OLAP permite ver los datos en función de muchas dimensiones, posibilitando a los usuarios expresar los datos de la forma que deseen. Entre las funcionalidades que puede ofrecer OLAP, se incluye la declaración de dimensiones, jerarquías, óptima indexación de los datos, definición de operaciones predefinidas de navegación en las dimensiones y de agrupación de medidas, por tanto entre los parámetros que definen estas funcionalidades se destacan [16] [9]:

- ✓ Seleccionar solo la información de un miembro en particular de una dimensión, o lo que es lo mismo, se trabaja con un subconjunto del total de los datos.

- ✓ Pasar la información del nivel anterior de la dimensión actual a una jerarquía definida, consolidando los datos del nivel actual y mostrando el valor consolidado, correspondiente al nivel inmediatamente superior.
- ✓ Realizar la operación inversa del Roll-up (operación con que se define el nivel de granularidad con la que se analizan los datos). Permite ver la información del nivel siguiente de la dimensión actual en una jerarquía definida.
- ✓ Cambiar la dimensión que está caracterizando los datos actualmente considerados.
- ✓ Visualizar la información de otro miembro del mismo nivel de la dimensión que se está evaluando. No detalla ni consolida la información, sino que cambia el miembro para el cual se están presentando los datos.
- ✓ Permitir la consulta de información del nivel inferior en la dimensión actual y la navegación por fuera del modelo multidimensional. La ejecución de esta operación depende de poder establecer el acceso al sistema fuente desde el OLAP.

Dentro de este sistema se pueden encontrar tres tipos de almacenamiento: el relacional, el multidimensional y el híbrido.

1.6.1 ROLAP

En Procesamiento Analítico en Línea Relacional (ROLAP) se accede a los datos almacenados en el MD para proporcionar los análisis OLAP, su principal premisa es extraer la información de bases de datos relacionales [17]. ROLAP utiliza una arquitectura de tres niveles. La base de datos relacional maneja el almacenamiento de los datos, el motor OLAP proporciona la funcionalidad analítica y alguna herramienta especializada es empleada para el nivel de presentación [18].

1.6.2 MOLAP

Un sistema MOLAP usa una Base de Datos Multidimensional (BDMD), en la que la información se almacena multidimensionalmente. MOLAP utiliza una arquitectura de dos niveles: la BDMD y el motor analítico. La BDMD es la encargada del manejo, acceso y obtención de los datos.

La información procedente de los sistemas transaccionales se carga en el sistema MOLAP. Una vez cargados los datos en la BDMD, se realiza una serie de cálculos para obtener datos agregados a través de las dimensiones del negocio, poblando la estructura de la BDMD; luego de llenar esta estructura, se generan índices y se emplean algoritmos de tablas de dispersión para mejorar los tiempos de accesos de las consultas [19].

1.6.3 HOLAP

Solución OLAP híbrida que combina el uso de las arquitecturas ROLAP y MOLAP. En una solución con HOLAP, los registros detallados (los volúmenes más grandes) se mantienen en la base de datos relacional, mientras que los agregados lo hacen en un almacén MOLAP independiente. Un sistema HOLAP resuelve el problema de dispersión, dejando los datos más granulares (menos agregados) en la base de datos relacional, pero almacena los agregados en un formato multidimensional, minimizando así la presencia de celdas vacías [19].

1.6.4 Valoración general

A continuación, se presenta una tabla comparativa de los tipos de almacenamiento antes mencionados:

Tabla 1. Comparativa entre los sistemas de almacenamiento OLAP [20].

Criterios	Tipo de almacenamiento		
	MOLAP	ROLAP	HOLAP
Almacenamiento de las agregaciones	Modelo multidimensional	Base de datos relacionales	Modelo multidimensional
Almacenamiento de datos	Modelo multidimensional	Base de datos relacionales	Base de datos relacionales
Facilidad de creación	Sencillo	Muy sencillo	Sencillo
Velocidad de respuesta	Buena	Regular o baja	Buena para consultas que posean agregaciones, Regular para datos de bajo nivel
Escalabilidad	Problemas de escalabilidad	Son más escalables	Escalable
Recomendado para	Cubo con uso frecuente	Datos que no son frecuentemente usados	Si el cubo requiere una rápida respuesta

Se propone para la solución ROLAP, dado que está enfocada en el modelo relacional y en ese sentido ofrece ventajas de escalabilidad sobre los demás sistemas de almacenamiento. Además, ROLAP ahorra espacio de almacenamiento y es útil cuando se trabaja con amplios conjuntos de datos. Su ventaja principal reside en la posibilidad de utilizar una tecnología ampliamente extendida y utilizada para la gestión de datos, los sistemas relacionales.

1.7 Modelado dimensional

Se recurre al modelo dimensional cuando se diseñan mercados de datos, almacenes de datos y sistemas de inteligencia de negocios. Teniendo un fin analítico es necesario utilizar esta técnica de modelado que permiten registrar las relaciones existentes entre los datos para ofrecer una fuente de consulta que facilite el análisis y toma de decisiones. El modelo dimensional divide el mundo de los datos en hechos, dimensiones y medidas. [21]

Tabla de hechos

Todo AD tiene incluido una o varias tablas de hechos, eso depende en gran medida de su complejidad. Estas tablas tienen como atributos una o más medidas de un proceso organizacional, de acuerdo a los requisitos. Un registro contiene una medida expresada en números, como: cantidad, tiempo, dinero, sobre la cual se desea realizar una operación de agregación (promedio, conteo, suma) en función de una o más dimensiones. La granularidad es el nivel de detalle que posee cada registro de una tabla de hechos.[22]

Están compuestas por un gran número de filas en dependencia de los años de trabajo de un proyecto, y su capacidad de almacenar información está a un nivel elevadísimo de detalle, esto sobresale como una de sus principales características. No deben incluir datos que no procedan de los campos numéricos y los campos de índice, guardan relación entre los hechos con las entradas en las tablas de dimensiones correspondientes. Este tipo de tabla incluye atributos para calcular las medidas y parámetros para determinar la descomposición de los datos, estos se describen como [9]:

- ✓ Medidas: Una medida es un atributo (campo) de una tabla que se desea analizar, agrupando sus datos, usando los criterios de corte conocidos como dimensiones [7]. Las medidas más útiles para incluir en una tabla de hechos son aquellas medidas numéricas que pueden calcularse con la suma de varias cantidades de la tabla. Es decir, las medidas candidatas son los datos numéricos, pero no cada atributo numérico es una medida candidata. En consecuencia, por lo general los hechos a almacenar en una tabla de hechos van a ser casi siempre valores numéricos, enteros o reales.
- ✓ Granularidad: La granularidad significa especificar el nivel de detalle. La elección de la granularidad depende de los requisitos del negocio y lo que es posible a partir de los datos actuales [7]. La granularidad de la tabla de hechos representa el nivel más atómico por el cual se definen los datos.

Tabla de dimensiones

Estas acompañan a la tabla de hechos y determinan los parámetros (dimensiones) de los que dependen los hechos registrados en la tabla de hechos. Son elementos que contienen atributos (o campos) que se utilizan para restringir y agrupar los datos almacenados en una tabla de hechos cuando se realizan consultas sobre dichos datos en un entorno de AD o MD. Los atributos de las dimensiones sirven como fuente primaria de las restricciones de las consultas, agrupaciones y las etiquetas de los reportes [9].

Atributos:

Los atributos son criterios utilizados para analizar las medidas. Se basan, en los datos de referencia de las tablas de dimensiones, además son campos o criterios de análisis, pertenecientes a las tablas de dimensiones [23].

Las tablas de hechos y dimensiones poseen distintas formas de representación, basado en esquemas, estos se definen como [24]:

- Esquema en estrella: En el esquema en estrella la tabla de hechos es la única tabla que tiene múltiples uniones que la conectan con otras tablas. El resto de las tablas del esquema (tablas de dimensión) únicamente hacen uniones con esta tabla de hechos. Toda la información referente a una dimensión, se almacena en la misma tabla.

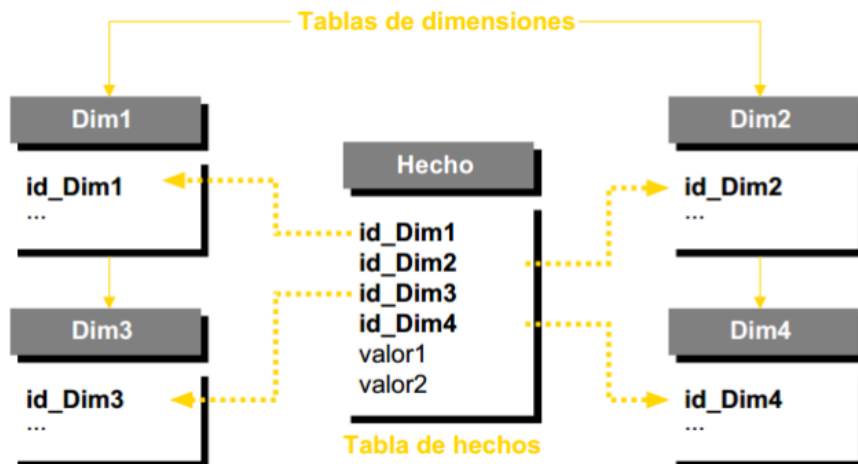


Figura 1. Esquema estrella

- Esquema copo de nieve: En cada dimensión se almacenan jerarquías de atributos o bien simplemente se separan atributos en otra entidad por razones de desempeño y mejor utilización del espacio. El uso más común de este esquema, es cuando las tablas de dimensiones son muy

grandes o complejas y es muy difícil representar los datos en un esquema estrella. En este esquema las tablas de dimensiones están normalizadas, con respecto al esquema estrella.

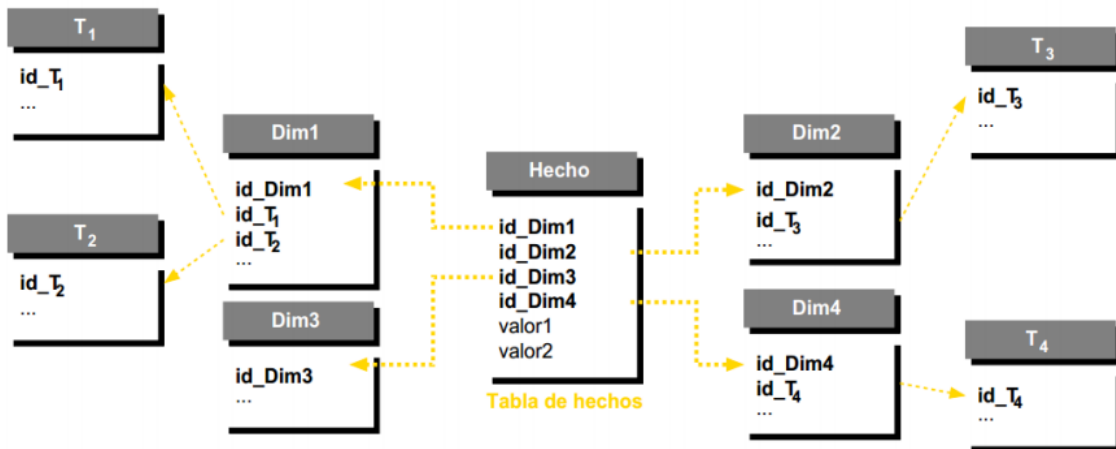


Figura 2. Esquema copo de nieve

- Esquema constelación de hechos: Para cada esquema estrella o esquema copo de nieve de un AD es posible construir un esquema de constelación de hechos. Este esquema es más complejo que las otras arquitecturas, debido a que contiene múltiples tablas de hechos. Con esta solución las tablas de dimensiones pueden estar compartidas entre más de una tabla de hechos.

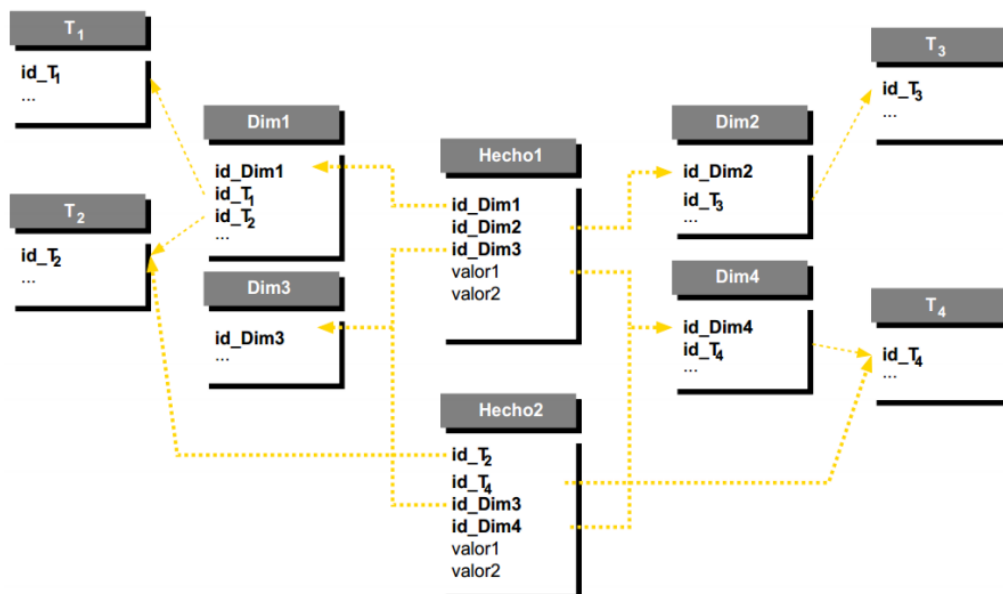


Figura 3. Esquema constelación de hechos

En la construcción de un MD se debe hacer una representación del modelo dimensional, en el diseño de este modelo se debe tener en cuenta uno de los esquemas descritos anteriormente. Esta técnica de

modelado permite registrar las relaciones existentes entre los datos para ofrecer una fuente de consulta que facilite el análisis y la toma de decisiones.

1.8 Metodología para el desarrollo del Mercado de datos

Con el desarrollo y avance en el nivel de conocimientos en los procesos de construcción de un MD, han surgido un conjunto de metodologías que agilizan y rigen los procesos de desarrollo de dichos sistemas. A continuación, se explica la metodología seleccionada, definida en la Universidad para el desarrollo de AD y MD.

1.8.1 Metodología para el desarrollo de Almacén de Datos propuesta por el Centro de Tecnologías de Gestión de Datos (DATEC)

Para el desarrollo del MD se define como metodología a utilizar: Metodología para el desarrollo de AD propuesta por DATEC, que toma como base el enfoque de Kimball y se ajusta a las condiciones y características de la producción del centro y la Universidad.

Se tomó como base la Metodología de Kimball por los siguientes elementos:

- 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 AD a través de la construcción de los MD departamentales, lo que constituye una estrategia buena y coincide con la división lógica de las empresas, entidades, organismos, entre otros.
- Es una metodología madura y reconocida por el resto de la comunidad dedicada al tema. Tiene bien definidas las etapas, actividades, artefactos y roles [4].

Durante el ciclo de vida de esta metodología se destacan las siguientes fases de desarrollo [4]:

- Estudio preliminar o planeación: se realiza el estudio de la entidad cliente, la planeación del proyecto, se definen los objetivos, el alcance preliminar, los costos estimados y otras actividades.
- Requisitos: se realiza en dos direcciones, una, mediante la identificación de las necesidades de información y reglas del negocio; y la otra con un levantamiento detallado de las fuentes de datos a integrar. Después se procede a la definición de los requerimientos.
- Arquitectura y diseño: se definen las estructuras de almacenamiento, se diseñan las reglas de extracción, transformación y carga, definiéndose la arquitectura de información que regirá el desarrollo de la solución.

- **Implementación:** se diseña físicamente el repositorio de datos, se crean las estructuras de almacenamiento, el área temporal de almacenamiento, se ejecutan las reglas de ETL y se configuran e implementan las herramientas de inteligencia de negocios para la obtención de los elementos que se acordaron con el cliente final.
- **Prueba:** se realizan las pruebas al sistema desde las Pruebas de Unidad hasta las de aceptación con el cliente final.
- **Despliegue:** se realiza un despliegue piloto en el cual se configuran los servidores y se instalan las herramientas y se carga una muestra de los datos para demostrar que el sistema funciona. Posterior a la aceptación del cliente se realiza la carga de los datos, así como la capacitación y transferencia tecnológica.
- **Soporte y mantenimiento:** tras la implantación de la solución se brindan los servicios de soporte en línea, vía telefónica, web u otras según el contrato firmado y las condiciones de soporte establecidas.
- **Gestión y administración del proyecto:** a lo largo del ciclo de vida se realizan actividades de control, gestión y chequeo del desarrollo, los gastos, las utilidades, los recursos y demás actividades por parte del grupo de dirección del proyecto.

Para la realización de la solución, no se transitará por todas las fases del ciclo de vida de la metodología. El trabajo será desarrollado hasta la fase de prueba, ya que el objetivo es desarrollar el MD de Energías Renovables sin llegar a desplegar la solución. A continuación, se describen las actividades a realizar por cada una de las fases de la metodología de DATEC.

1.8.2 Descripción de las actividades en cada fase

Las actividades se agruparon por cada una de las fases de la metodología por la que se transitará. Cada actividad está compuesta por un grupo de tareas que detallan las acciones que deben ejecutarse para su adecuado desarrollo [4]:

Estudio preliminar y planeación:

- Examinar la situación actual de la organización, sus perspectivas respecto al cumplimiento de su misión, sus objetivos y actividades.
- Identificar las necesidades del negocio.
- Definir el equipo de trabajo.

Requisitos:

- Identificar requisitos de información.

- Identificar requisitos no funcionales.

Arquitectura y diseño:

- Identificar hechos, dimensiones y medidas.
- Diseñar modelo dimensional.
- Diseñar el modelo de integración de datos.
- Identificar reportes candidatos.
- Diseñar el cubo OLAP.

Implementación:

- Implementación del MD.
- Realizar la carga histórica para las dimensiones y el hecho.
- Configurar los niveles de acceso a la información.
- Desarrollar la solución de BI.

Pruebas:

- Realizar las pruebas al MD.

1.9 Conclusiones del capítulo

En este capítulo se realizó un estudio teórico y conceptual sobre los AD y MD. Sobre la base de dicho análisis se decidió realizar un MD de Energías Renovables, utilizando para el desarrollo la metodología propuesta por DATEC. Además, se definen las actividades que se realizan en las diferentes fases de desarrollo y que garantizan la calidad del proceso de desarrollo.

Capítulo 2: Análisis y diseño

2.1 Introducción

En este capítulo se realiza el análisis y diseño del MD sobre energías renovables a partir de los datos publicados por la ONEI y por IRENA sobre estas energías. Para la construcción del MD se tienen en cuenta las necesidades de información del cliente, a partir de las cuales se definen y especifican los requisitos de información, los requisitos no funcionales y otros artefactos según la metodología definida. Además, se definen las herramientas para la elaboración de la solución, dejando plasmados los elementos por los cuales se seleccionan.

2.2 Necesidades del Negocio

A partir del encuentro de trabajo realizado entre el Centro de Gestión de la Información y Desarrollo de la Energía (CUBAENERGIA) y el grupo de investigación de Web Semántica, las principales dificultades que se identifican están centradas en no poder realizar análisis sobre la información que se publica sobre energías renovables. Las fuentes de energías renovables a tener cuenta para realizar análisis son las siguientes:

Tabla 2. Fuentes de energías renovables

Fuentes de energía	Eólica	Solar	Hidroeléctrica	Biomasa
Tipos de energía	Energía eólica marina Energía eólica terrestre	Energía solar fotovoltaica Energía solar termoeléctrica	Hidroeléctricas	Bagazo Paja de caña Serrín de madera Cáscara de arroz Desechos de café Cáscara de coco Desechos forestales Desechos agrícolas Leña

Los principales datos que se publican sobre las energías renovables hacen referencia a la capacidad de producción, cantidad producida, consumo físico de la biomasa y la cantidad de dispositivos generadores instalados en Cuba. A partir de estos datos se necesita conocer el ahorro que significaría en petróleo la utilización de estas energías, el cual se representa en la unidad de medida: equivalente en toneladas de petróleo.

Luego de estudiar las necesidades del negocio se hace necesario conocer los detalles de las fuentes de datos, por lo cual se procede a definir el registro de sistemas fuentes. Los registros de sistema fuente describen las características de las fuentes de datos teniendo en cuenta: el tipo de fuente (externa o interna), los formatos en que se encuentran los datos, las condiciones de acceso, la periodicidad de actualización, la cantidad de datos que manejan, entre otros aspectos de interés para definir los procesos de integración de datos. [4]

Tabla 3. Estudio de las fuentes de datos

Fuentes de datos	Tipo de fuente	Formato en que se encuentran los datos	Periodicidad de actualización	Cantidad de documentos Excel a procesar
ONEI	Externa	Excel	anual	40
IRENA	Externa	Excel	anual	14

2.3 Requisitos

En la fase de Requisitos, definida en la propuesta de metodología para el desarrollo de AD en DATEC para la construcción de un MD, se identifican las necesidades de información de la organización, las características y cualidades del sistema. En la definición inicial de una solución de MD se deben de identificar tres tipos de requisitos: los de información, los funcionales y los no funcionales. Para la realización del MD de Energías Renovables no se definieron requisitos funcionales ya que para realizar la extracción, transformación y carga se utilizó la herramienta Pentaho Data Integration v6.0 y para la gestión de usuarios, reportes, roles y visualización de la información, se utilizaron las funcionalidades que brinda la herramienta Pentaho BI Server v6.0.

2.3.1 Requisitos de información

Los requisitos de información describen la información y los datos que el sistema debe almacenar para satisfacer a los clientes. Esto se define a partir de las necesidades de información identificadas, que permitan realizar el análisis del comportamiento de dicha información. Los requisitos de información identificados durante el proceso de análisis se muestran a continuación.

A nivel nacional:

RI1. Determinar el ahorro de petróleo por dispositivo generador y provincia.

RI2. Obtener la cantidad de dispositivos generadores de energía renovable utilizados por año.

RI3. Obtener el consumo físico de biomasa empleada como combustible por año.

RI4. Determinar la cantidad de energía producida equivalente en petróleo de la biomasa por provincias.

RI5. Obtener el equivalente en toneladas de petróleo, la producción de energía por dispositivo generador y año.

RI6. Obtener las cinco provincias que más ahorran en petróleo.

A nivel internacional:

RI7. Obtener la capacidad de producción de energía renovable por fuentes renovables y año.

RI8. Obtener la cantidad de producción de energías renovables dado un país por fuentes y año.

RI9. Obtener los diez países que más han producido energía renovable.

RI10. Obtener la cantidad producida de energía renovable por país y región.

RI11. Obtener la cantidad que representa en ahorro de petróleo, la producción de energía de fuentes renovables por año, regiones y países.

RI12. Obtener la fuente de energía renovable con mayor producción de energía.

2.3.2 Requisitos no funcionales

Los requisitos no funcionales son aquellos requerimientos que no se refieren directamente a las funcionalidades específicas que proporciona el sistema, sino a las propiedades de este como: la fiabilidad, el tiempo de respuesta y la capacidad de almacenamiento [25]. A continuación, se listan los que debe cumplir el sistema en proposición.

Software:

RNF1: Las estaciones de trabajo clientes deben contar con un navegador web (en caso de utilizar el navegador Firefox, una versión posterior a la 2.0). El servidor de aplicaciones debe contar con las herramientas: Pentaho BI Server 6.0, Java Virtual Machine 7.0 o superior, Apache Tomcat 6.0 y Un navegador web. El servidor de base de datos debe contar con las herramientas: PostgreSQL v9.4 como Sistema Gestor de Base de Datos y PgAdmin 1.20 como Administrador de Base de Datos.

Hardware:

RNF2: Proporcionar características mínimas de hardware a las estaciones de trabajo clientes y servidores: las estaciones de trabajo deben contar con un mínimo de 1 Gb de memoria RAM. Los servidores para lograr una explotación aceptable del sistema deben contar con los requerimientos de hardware: 8 Gb RAM y 80 Gb de disco duro.

Usabilidad:

RNF3: Agilizar el acceso a los reportes del almacén de datos mediante la distribución de la información por áreas de análisis y libros de trabajo: El usuario podrá acceder de manera rápida a la información que solicita, la cual se encontrará ubicada dentro del libro de trabajo y área correspondiente, de acuerdo al objetivo de su solicitud.

Interfaz:

RNF4: Acceder al sistema: El usuario deberá acceder a la aplicación utilizando un navegador web.

Eficiencia:

RNF5: Los reportes y estadísticas que se obtendrán deben ser precisos y reales.

Restricciones en el diseño y la implementación

RNF6: Para la programación en el MD se realizarán las consultas en MDX.

RNF7: De la suite Pentaho, se utilizarán las herramientas Schema Workbench 3.11.0 y Pentaho BI Server 6.0. Para el uso de las herramientas anteriores se requiere la instalación de la máquina virtual de java Java Virtual Machine 7.0 o una versión superior.

2.4 Casos de uso de información

Los casos de uso de información (CUI) agrupan un conjunto de requisitos de información teniendo en cuenta los conceptos del negocio que manejan, fundamentalmente por tema de análisis. En la presente investigación se identificó un tema de análisis, generando dos agrupaciones de CUI.

La herramienta que se utiliza para el modelado es Visual Paradigm v5.0, esta es una herramienta CASE⁴ de modelado visual que facilita la construcción de artefactos en un proceso de desarrollo de software mediante el Lenguaje de Modelado Unificado (UML). Soporta una amplia gestión de casos de uso y diseño de base de datos relacionales y proporciona medidas más eficaces en el análisis y diseño

⁴ CASE: por sus siglas en inglés, Computer Aided Software Engineering.

de sistemas. Algunas de las funcionalidades que brinda la herramienta son la generación de código y de base de datos a partir de los diagramas UML realizados, la realización de ingeniería inversa, generación automática de informes en formato PDF, Word o HTML [26].

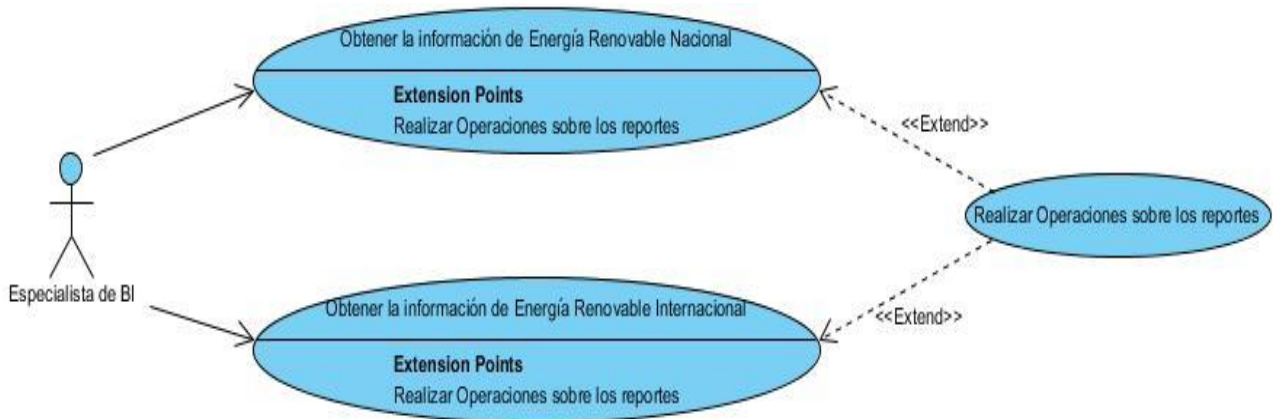


Figura 4. Diagrama de casos de uso de información

CUI1 Obtener la información de la energía renovable nacional: visualiza los reportes de los requisitos de información de la energía renovable nacional.

CUI2 Obtener la información de la energía renovable internacional: visualiza los reportes de los requisitos de información de la energía renovable internacional.

Tabla 4. CUI1 Obtener la información de energía renovable nacional

Objetivo	Mostrar toda la información referente a la energía renovable.
Actores	Especialista de BI: (Inicia) seleccionando el Área de análisis.
Resumen	El caso de uso se inicia cuando el actor desea hacer un análisis de la información relacionada con las energías renovables nacional. El actor selecciona el reporte que desea ver, el sistema muestra la información contenida y a partir de la información mostrada el actor realiza los cambios deseados al reporte. Luego de esto finaliza el caso de uso.
Complejidad	Alta
Prioridad	Media
Precondiciones	El usuario se autenticó correctamente. Los datos correspondientes fueron cargados en el MD.

	Los reportes relacionados con las energías renovables fueron creados.																	
Postcondiciones	Los reportes correspondientes al caso de uso fueron consultados.																	
Flujo de eventos																		
Flujo básico Presentar información referente a la capacidad, producción y dispositivos de energías renovables																		
	Actor		Sistema															
1.	Selecciona el Área de Análisis (A.A) de Energía Renovable.		Muestra el A.A de Energía Renovable.															
2.	Selecciona el libro de trabajo Energía Renovable.		Muestra el contenido en el libro de trabajo.															
3.	Selecciona en el libro de trabajo Análisis nacional.		Muestra el contenido en el libro de trabajo del análisis seleccionado.															
4.	Selecciona el reporte que desea analizar.		Muestra la información contenida en el reporte seleccionado y brinda opciones al actor para hacer cambios al reporte durante su análisis. Finaliza el CU.															
Prototipo de Interfaz de usuario																		
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="3"></th> <th colspan="2">Medidas</th> </tr> <tr> <th>Año</th> <th>Provincia</th> <th>Fuente de Energía Renovable</th> <th>● Consumo físico</th> <th>● Equivalente_en_petróleo</th> </tr> </thead> <tbody> <tr> <td style="text-align: center;">+ Total</td> <td style="text-align: center;">+ Total</td> <td style="text-align: center;">+ Total</td> <td style="text-align: center;">38.162.722</td> <td style="text-align: center;">8.644.818</td> </tr> </tbody> </table>							Medidas		Año	Provincia	Fuente de Energía Renovable	● Consumo físico	● Equivalente_en_petróleo	+ Total	+ Total	+ Total	38.162.722	8.644.818
			Medidas															
Año	Provincia	Fuente de Energía Renovable	● Consumo físico	● Equivalente_en_petróleo														
+ Total	+ Total	+ Total	38.162.722	8.644.818														
Relaciones	CU incluidos	N/A																
	CU extendidos	Realizar operaciones sobre los reportes.																
Requisitos no funcionales																		
Asuntos pendientes	N/A																	

Tabla 5. CUI2 Obtener la información de energía renovable internacional

Objetivo	Mostrar toda la información referente a la energía renovable.	
Actores	Especialista de BI: (Inicia) seleccionando el Área de análisis.	
Resumen	El caso de uso se inicia cuando el actor desea hacer un análisis de la información relacionada con las energías renovables internacional. El actor selecciona el reporte que desea ver, el sistema muestra la información contenida y a partir de la información mostrada el actor realiza los cambios deseados al reporte. Luego de esto finaliza el caso de uso.	
Complejidad	Alta	
Prioridad	Media	
Precondiciones	El usuario se autenticó correctamente. Los datos correspondientes fueron cargados en el MD. Los reportes relacionados con las energías renovables fueron creados.	
Postcondiciones	Los reportes correspondientes al caso de uso fueron consultados.	
Flujo de eventos		
Flujo básico Presentar información referente a la capacidad, producción y dispositivos de energías renovables		
	Actor	Sistema
5.	Selecciona el Área de Análisis (A.A) de Energía Renovable.	Muestra el A.A de Energía Renovable.
6.	Selecciona el libro de trabajo Energía Renovable.	Muestra el contenido en el libro de trabajo.
7.	Selecciona en el libro de trabajo el Análisis internacional.	Muestra el contenido en el libro de trabajo del análisis seleccionado.
8.	Selecciona el reporte que desea analizar.	Muestra la información contenida en el reporte seleccionado y brinda opciones al actor para hacer cambios al reporte durante su análisis. Finaliza el CU.
Prototipo de Interfaz de usuario		

País	Año	Fuente de Energía Renovable	Medidas		
			• Capacidad de producción	• Equivalente en toneladas de petróleo	• Cantidad producida
+ Total	+ Total	+ Total	15.036.556	3.644.716.120	42.380.420
Relaciones		CU incluidos	N/A		
		CU extendidos	Realizar operaciones sobre los reportes.		
Requisitos no funcionales					
Asuntos pendientes		N/A			

2.5 Arquitectura de un mercado de datos

Una vez identificados los requisitos de información, se procede a definir la arquitectura del MD. Esta describe todo el flujo de datos, desde que son extraídos de los sistemas fuentes, hasta su preparación para ser utilizados por los clientes del negocio. Presenta cuatro niveles:

- ✓ **Fuente de datos:** se refiere al origen de los datos, son los archivos de extensión .xls que contienen la información.
- ✓ **Subsistema de integración de datos:** Es donde se realizan los procesos ETL. Es el nivel encargado de conectarse a las fuentes de datos y extraer la información, realizar las transformaciones necesarias y finalmente carga la información en las fuentes destinos.
- ✓ **Subsistema de almacenamiento:** base de datos relacional que contiene las tablas de dimensiones y hechos cargadas a través de los procesos de ETL para que puedan ser consultadas.
- ✓ **Subsistema de visualización:** comprende las interfaces orientadas a usuario que extraen información y le facilitan la toma de decisiones por medio de herramientas de Procesamiento Analítico en Línea (OLAP) y de técnicas de Inteligencia de Negocio.



Figura 5. Arquitectura del mercado de datos

2.6 Diseño de la solución

El diseño parte del análisis realizado sobre la base de los requisitos definidos anteriormente y que cumple con los objetivos del sistema. Este debe ser suficientemente claro para que el sistema pueda ser implementado sin ambigüedades.

2.6.1 Diseño del subsistema de almacenamiento

Para el almacenamiento de los datos es necesario un Sistema Gestor de Base de Datos (SGBD), que es un tipo de software específico, dedicado a servir de interfaz entre la base de datos, el usuario y las aplicaciones que lo utilizan [5]. En esta propuesta de solución se utilizará PostgreSQL v9.4.

Para el diseño del subsistema de almacenamiento es necesario identificar las dimensiones y los hechos con sus medidas asociadas, así como las relaciones existentes entre los hechos y las dimensiones.

Dimensiones

Las tablas de dimensiones son elementos que contienen atributos utilizados para restringir y agrupar los datos que participan en el análisis. Las dimensiones poseen entre sus características principales la definición de jerarquías entre sus atributos, que tienen como objetivo plasmar explícitamente la forma

en que se puede consolidar. Son las compañeras integrales de las tablas de hechos, ellas contienen la descripción textual del negocio [11].

Las dimensiones identificadas se describen a continuación:

1. Dimensión tipo de energía renovable (*dim_tipos_energias_renovables*): Guarda la información referente a la fuente renovable y a los tipos de energía renovable, permitiendo conocer los atributos: `codigo_tipo_energia_renovable`, `nombre_tipo_energia_renovable`, `codigo_tipo_fuente_energia` y `nombre_tipo_fuente_energia`.






dimension_tipos_energias_renovables	
 pk_tipo_energia_renovable	int4
 <code>codigo_tipo_energia_renovable</code>	int4
 <code>nombre_tipo_energia_renovable</code>	varchar(50)
 <code>codigo_tipo_fuente_energia</code>	int4
 <code>nombre_tipo_fuente_energia</code>	varchar(50)

Figura 6. Dimensión tipos de energías renovables

2. Dimensión dispositivo generador (*dim_dispositivo_generador*): Contiene los datos referentes a los dispositivos generadores, permitiendo conocer el atributo: `nombre_dispositivo_generador` y el `codigo_dispositivo_generador`.




dimension_dispositivo_generador	
 pk_dimension_dispositivo_generador	int4
 <code>codigo_dispositivo_generador</code>	int4
 <code>nombre_dispositivo_generador</code>	varchar(50)

Figura 7. Dimensión dispositivos generadores

3. Dimensión país (*dim_pais*): Contiene los datos referentes al país, así como a la región que pertenece y las provincias de Cuba, permitiendo conocer los atributos: `codigo_pais`, `nombre_pais`, `codigo_region`, `nombre_region`, `codigo_provincia` y `nombre_provincia`.

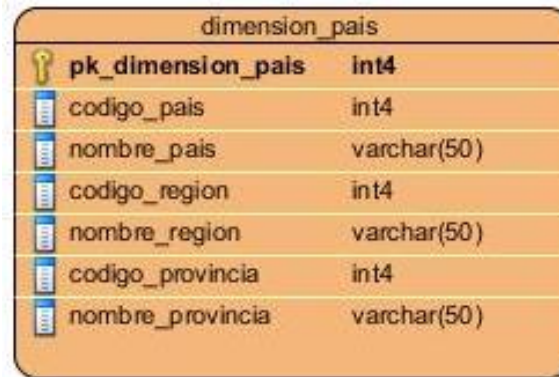


Figura 8. Dimensión país

- Dimensión año (dim_anno): Esta dimensión es una de las comunes e importantes en el diseño de MD, debido a que define el tiempo para enmarcar la información almacenada y organizarla atendiendo al momento en que fue captada. La misma permite conocer los atributos: numero_anno y codigo_anno.

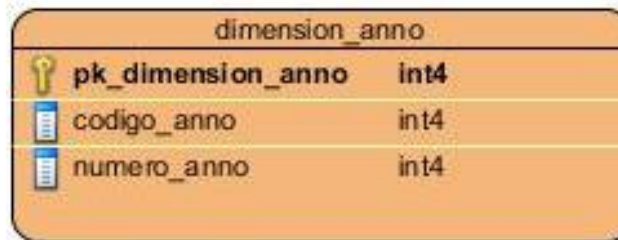


Figura 9. Dimensión año

Hecho

La tabla de hecho contiene los hechos que serán utilizados por los especialistas de BI para apoyar el proceso de toma de decisiones y entre su contenido están los datos cuantitativos. Esta se encuentra rodeada de las tablas de dimensiones. Para el desarrollo del MD se identificó un hecho:

- Hecho producción de energías renovables (hech_produccion_energias_renovables): Contiene la capacidad de producción, la cantidad producida y la cantidad de dispositivos generadores en cuba. Así como el consumo físico y el equivalente en toneladas de petróleo, permitiendo conocer los atributos capacidad_producción, cant_producida, cant_dispositivos_generadores, consumo_fisico y equivalente_toneladas_petroleo.











hecho_produccion_energias_renovables	
 pk_hecho_energia_renovable	int4
 fk_dimension_pais	int4
 fk_dimension_anno	int4
 fk_tipo_energia_renovable	int4
 fk_dimension_dispositivo_generador	int4
 capacidad_produccion	int4
 cant_producida	int4
 cant_dispositivos_generadores	int4
 consumo_fisico	int4
 equivalente_toneladas_petroleo	int4

Figura 10. Hecho producción de energías renovables

Matriz dimensional

La matriz dimensional representa la relación que existe entre la tabla de hecho y las tablas de dimensiones con las que se relaciona. Seguidamente, se muestra la matriz correspondiente al MD de Energías Renovables, donde se reflejan las relaciones entre el hecho y las dimensiones explicadas anteriormente.

Tabla 6. Matriz dimensional

Hecho	Dimensiones			
	Tipos de energías renovables	País	Dispositivos generadores	Año
Producción de energías renovables	X	X	X	X

Modelo de datos

Una vez definido dentro del negocio las dimensiones y medidas se procede a la estructuración del modelo dimensional. A continuación, se muestra el modelo de datos diseñado para el desarrollo del MD, donde se evidencia la representación del esquema estrella, atendiendo a que existe un hecho y varias dimensiones.

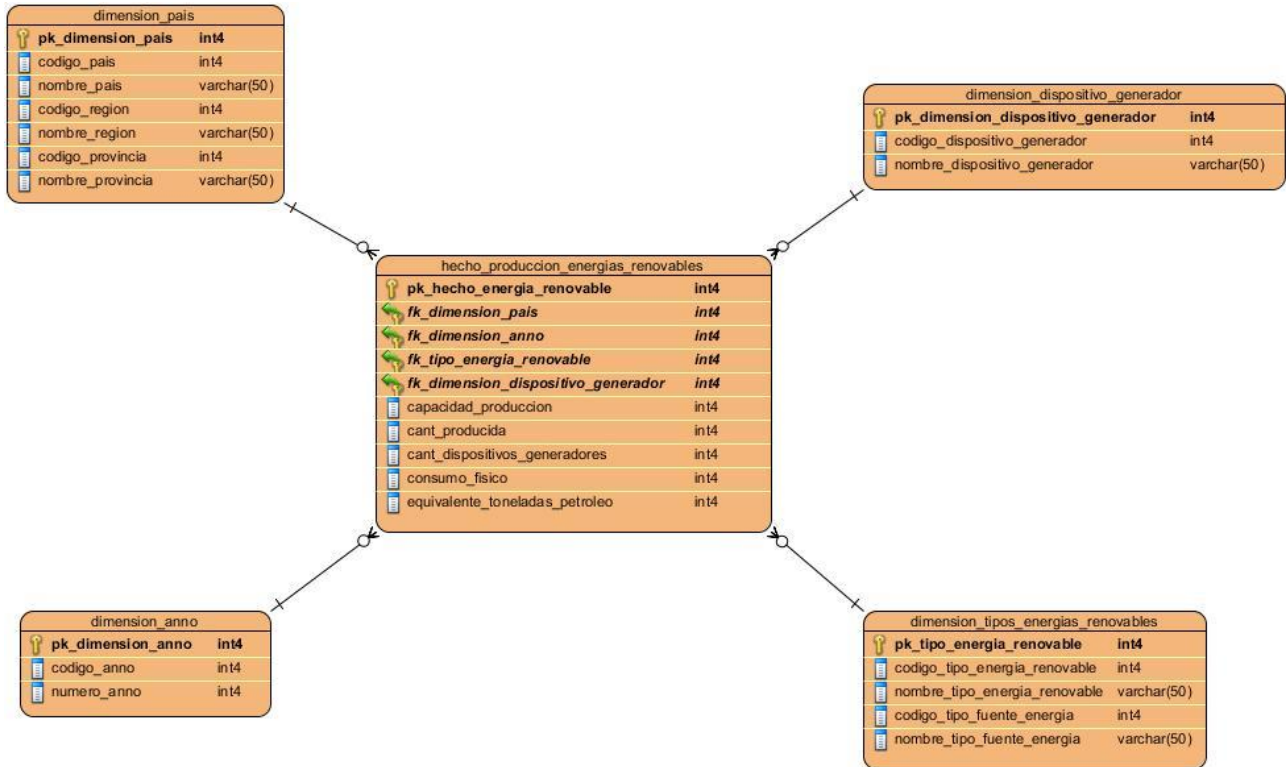


Figura 11. Modelo de datos del AD

2.6.2 Diseño del subsistema de integración

Para los procesos ETL, es preciso contar con herramientas que permitan reducir tiempo y costos. El desarrollo y la diversificación de las herramientas para los procesos de ETL, actualmente son crecientes, y se refleja en la amplia variedad de herramientas tanto comerciales como de código abierto [27]. La herramienta a utilizar para la integración de los datos es Pentaho Data Integration v6.0 para realizar la ETL. Es una herramienta multiplataforma que permite extraer la información de las diferentes fuentes, transformar la información a través de un modelo dimensional y cargar los resultados de la transformación en una base de datos tipo AD, a partir del entorno gráfico de desarrollo [27].

Una vez que se conoce la estructura, contenido y fiabilidad de los datos, se procede a realizar el diseño de las transformaciones. Durante la implementación estas transformaciones pueden sufrir cambios debido a disímiles situaciones que ocurren con los datos y para contrarrestar esto se llevan a cabo varias estrategias.

A continuación, se muestra el procedimiento que se lleva a cabo para cargar los datos en el MD de energía renovable:

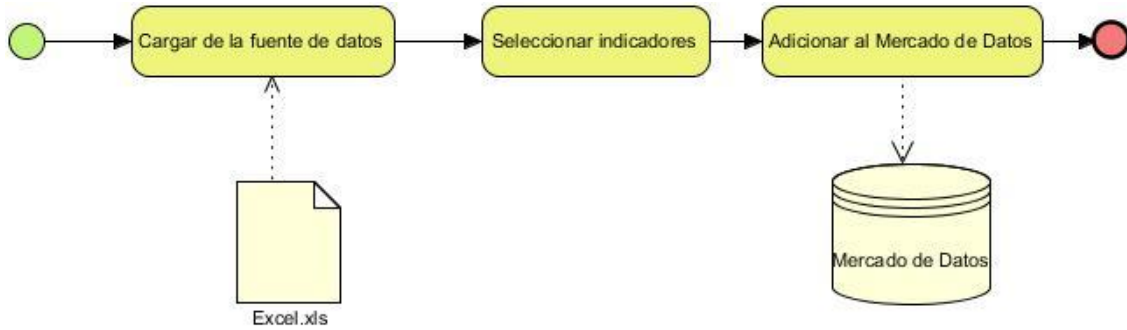


Figura 12. Diseño de la transformación para la carga de los datos

En el caso de las tablas de hechos, se diseña la siguiente transformación que permitirá cargar los datos hacia el MD Energía Renovable:

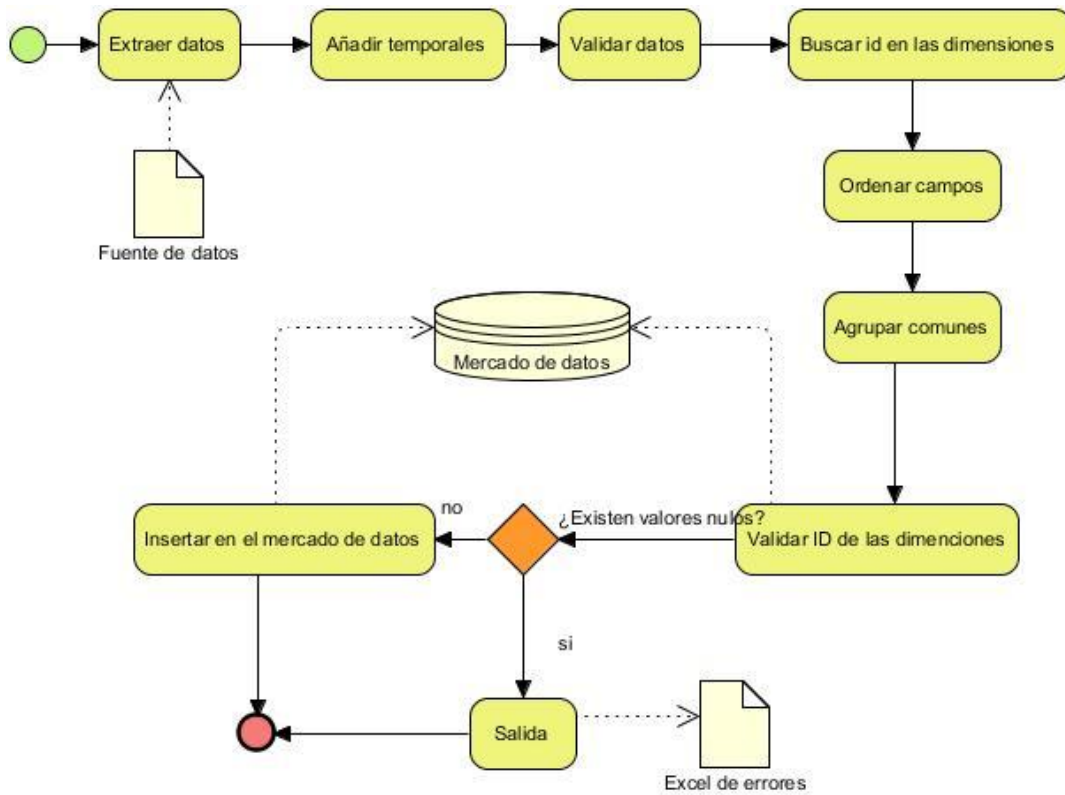


Figura 13. Diseño de la transformación para la carga de los datos en los hechos

2.6.3 Diseño del subsistema de visualización

Para realizar el diseño del subsistema de visualización es necesario diseñar los reportes que responden a las necesidades de información del cliente y los cubos OLAP. Esta actividad permite diseñar los cubos OLAP para el análisis multidimensional de la información. Se definen los reportes que serán implementados como parte de la solución.

Diseño de los cubos OLAP

En el proceso de creación de una estructura OLAP se debe tener presente que lo que se hace es mapear el diseño de la base de datos (tablas de hecho y dimensiones) con el diseño, de forma que es posible crear un esquema con la misma cantidad de elementos. En este proceso se consideran la tabla de hecho y las tablas de dimensiones. En el MD de Energías Renovables se tienen un cubo multidimensional, para representar la tabla de hecho, cuatro dimensiones y cinco medidas, representado como se muestra en la Figura 14.



Figura 14. Cubo producción de energía renovable

2.7 Política de respaldo y recuperación

Con el objetivo de garantizar la persistencia de la información, se establece una política de respaldo y de recuperación de los datos.

Periodicidad de las salvas: las salvas de toda la información contenida en la base de datos se realizarán con una periodicidad anual. Se debe verificar la existencia de una copia de toda la información almacenada. Las tablas involucradas en este proceso son: producción de energía renovable, producción de energía renovable nacional y producción de energía renovable nacional por dispositivo, en el proceso de análisis con las tablas de dimensiones asociadas.

Backups existentes: cada año se realizarán salvas a fin de año. Se debe realizar una copia de los datos en otros medios de almacenamiento, dígame memorias USB, discos duros externos o DVD.

2.8 Esquema de seguridad

Para evitar el daño de los datos y de las funcionalidades de una aplicación informática es necesario el desarrollo de un esquema de seguridad estable que garantice la seguridad y protección de los datos.

Esta seguridad se rige principalmente por los roles y permisos que los usuarios poseen para la interacción con el sistema.

2.8.1 Seguridad en la base de datos

Como complemento del esquema de seguridad mencionado anteriormente se definieron los roles mencionados a continuación para el acceso a la base de datos.

Tabla 7. Permisos de los usuarios en la base datos

Usuarios	Permisos
Administrador	Tiene total acceso a todas las funciones de la base de datos

2.8.2 Seguridad en la aplicación

Para el acceso de los usuarios al servidor BI de la Suit del Pentaho se definen los siguientes roles para lograr una mayor organización a la hora de acceder a la aplicación.

Tabla 8. Permisos de los usuarios en la aplicación

Usuarios	Permisos
Administrador	Tiene total acceso al sistema
Especialista de BI	Tiene acceso de solo lectura al MD

2.9 Conclusiones del capítulo

En este capítulo se abarcaron los principales elementos con respecto a las etapas de análisis y diseño de la solución. Se lograron identificar los requisitos de información y los requisitos no funcionales dando respuesta a las necesidades del cliente. El diseño del subsistema de almacenamiento logró generar el modelo de datos a partir de la definición de la tabla de hechos y las tablas de dimensiones, con el diseño del subsistema de integración se definió una estructura general para los procesos ETL y con el diseño del subsistema de visualización se generó el cubo OLAP para el análisis multidimensional de los datos.

Capítulo 3: Implementación y validación

3.1 Introducción

Una vez realizado el análisis y diseño del MD de Energías Renovables y teniendo como guía la metodología utilizada, se procede a realizar la implementación de los subsistemas y el proceso de validación. Esta nueva etapa comprende la implementación de cada uno de los subsistemas que conforman el MD: subsistemas de almacenamiento, integración y visualización. Una vez concluido el proceso de construcción del MD de Energías Renovables, se hace necesaria la validación de la solución mediante la aplicación de pruebas.

3.2 Implementación del subsistema de almacenamiento

Para la implementación del subsistema de almacenamiento fue necesaria la utilización de un gestor de base de datos. El SGBD utilizado fue PostgreSQL v9.4 que permitió representar el diseño de la solución, mediante la creación de las tablas con sus respectivos atributos y relaciones.

El desarrollo exitoso de un modelo físico es un aspecto a tener en cuenta dentro de la construcción de un MD. El punto de partida para llevar a cabo este modelo es el modelo dimensional elaborado en el capítulo anterior. A continuación, se describe lo relacionado con la implementación del modelo de datos para el Mercado de datos de Energías Renovables, describiendo la estructura de la base de datos conformada por los esquemas, tablas y secuencias.

3.2.1 Esquemas y tablas

Los esquemas y las tablas permiten brindar una idea general de la base de datos, así como analizar y profundizar los contenidos básicos de la misma. Para la solución propuesta se cuenta con cuatro tablas de dimensiones identificadas y una tabla de hechos. A continuación, se muestra la estructura de datos en el PostgreSQL.

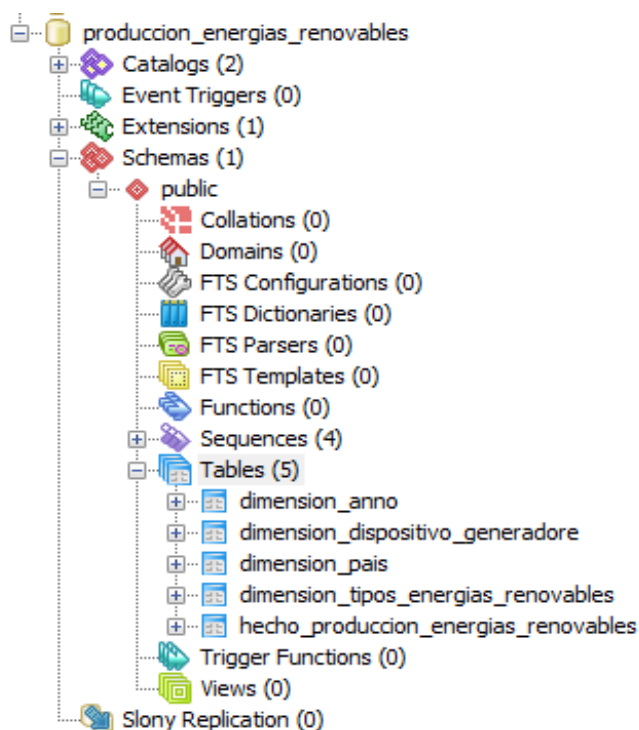


Figura 15. Estructura de datos

3.2.2 Restricciones y secuencias

Las restricciones son condiciones que se le aplican a una base de datos para que cumpla con ciertos parámetros. Pueden ser creadas automáticamente al definir una tabla (en el caso de las llaves primarias) o ser introducidas por el programador de la base de datos cuando se busca algo específico, por ejemplo, que un campo tenga una longitud determinada o tenga la estructura de un ciclo. Existen cuatro tipos de restricciones que son las más comunes [28]:

- Clave foránea: son las que referencian una clave de otra tabla, se usan para relacionar tablas diferentes.
- Duplicidad: implica que no debe haber dos valores iguales en la misma columna.
- Clave primaria: son valores que deben cumplir con un conjunto de restricciones: no tener valores nulos, ser únicos para cada tupla y ser necesarios.
- Valor no nulo: no debe existir ninguna casilla de la columna que esté vacía.

En el MD de Energías Renovables se utiliza el patrón de claves subrogadas, en cada tabla de dimensiones el tipo de clave es primaria y para la tabla de hecho, se tiene una clave primaria y el conjunto de claves foráneas provenientes de cada una de las dimensiones. Las secuencias son atributos que se incrementan secuencialmente durante el ingreso de los datos. En el presente trabajo se definieron cinco llaves primarias y cinco secuencias.

3.3 Implementación del subsistema de integración

Una vez definidas las fuentes de datos se realizan las transformaciones para identificar y corregir los problemas existentes en estas. Para ello se realiza la limpieza de los datos que permite llenar valores ausentes y reparar errores. Una vez que los datos son transformados se cargan, poblando las dimensiones y el hecho que conforma la estructura del subsistema de almacenamiento del Mercado de datos de Energías Renovables.

3.3.1 Implementación de las transformaciones

Las transformaciones constituyen un elemento básico dentro de la implementación del proceso de ETL. Una transformación está compuesta por pasos, que constituyen el elemento más pequeño de la transformación y se encuentran unidos a través de saltos. A continuación, se muestran algunas de las transformaciones realizadas para desarrollar la solución:

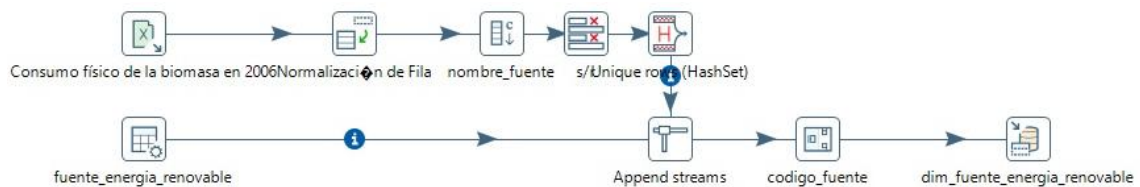


Figura 16. Transformación para la carga de la dimensión dimension_tipos_energias_renovables

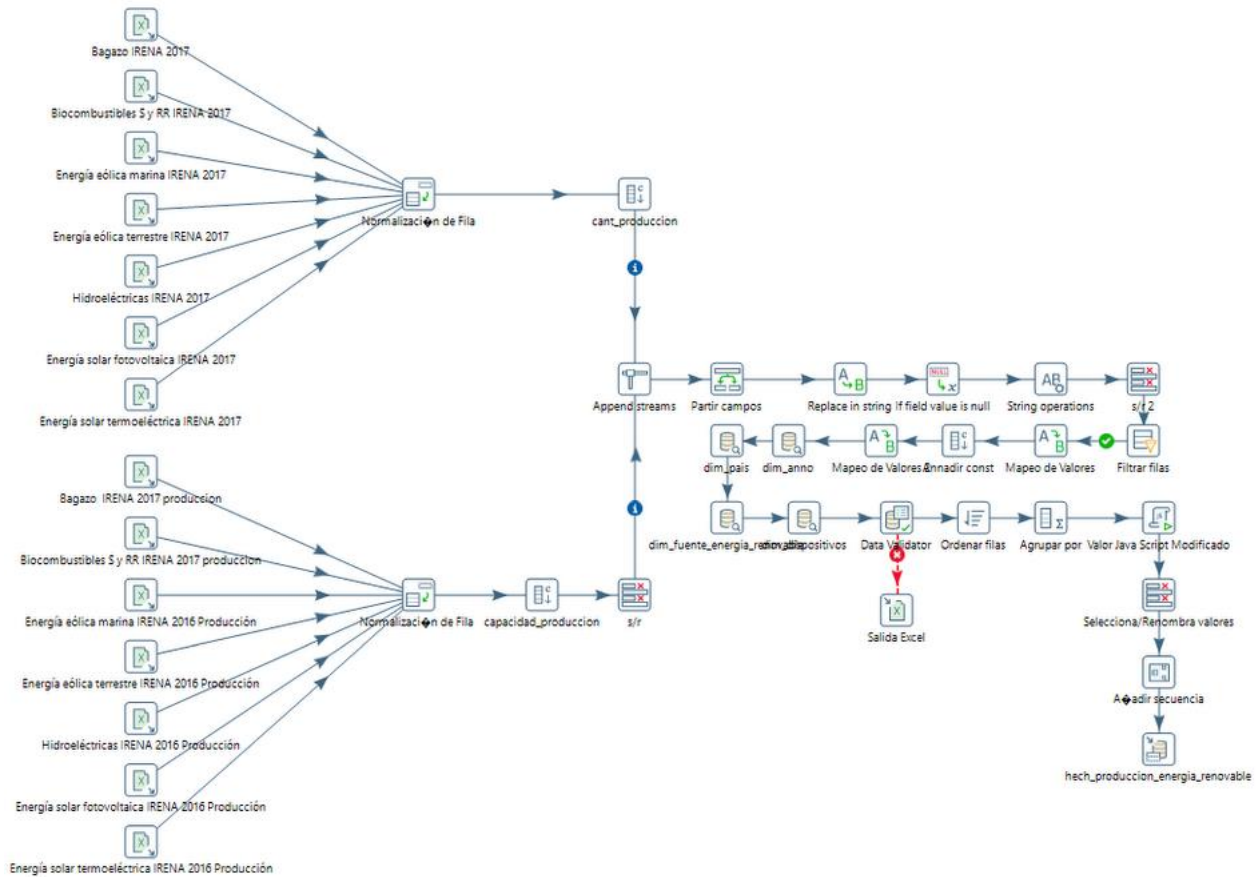


Figura 17. Transformación para la carga del hecho `hech_produccion_energias_renovables`

3.4 Implementación del subsistema de visualización

Una vez realizada la carga de los datos y como parte de la implementación del tercer y último subsistema del Mercado de datos de Energías Renovables, se realizó un cubo OLAP y los reportes candidatos, así como la configuración del control de acceso al sistema.

3.4.1 Cubos de datos

En el sistema se define un cubo OLAP, que abarca todas las dimensiones descritas anteriormente. Dicho cubo contiene las medidas necesarias para regir los hechos numéricos en esta estructura y según los objetivos estratégicos a cumplir.

El cubo de datos se desarrolla utilizando la herramienta Schema Workbench v3.11. El tipo de almacenamiento definido para este cubo es ROLAP que es el que utiliza la herramienta mencionada anteriormente. El primer paso para la creación del cubo es la definición del mismo y donde se deciden cuáles son los atributos que son necesarios para realizar el análisis. Como segundo paso se especifica la estructura de las dimensiones diseñando los campos calculables, las medidas y se crean todas las

tablas relacionales requeridas. Para poder realizar el análisis del cubo OLAP, debe publicarse dentro de la plataforma Pentaho. A continuación, se muestran las interfaces al publicar el cubo en el Pentaho BI server:

Países	Años	Dispositivos generadores	Energías renovables	Medidas				
				• Cantidad de dispositivos generadores	• Ahorro en toneladas de petróleo	• Consumo físico	• Capacidad de producción de energía (MW/h)	• Cantidad de energía producida (GW/h)
Total	Total	Total	Total	116.664	3.832.324.358	38.162.722	15.006.742	44.461.390

Figura 18. Interfaz de la vista de análisis de Energías renovables

Países	Años	Dispositivos generadores	Energías renovables	Medidas						
				• Cantidad de dispositivos generadores	• Ahorro en toneladas de petróleo	• Consumo físico	• Capacidad de producción de energía (MW/h)	• Cantidad de energía producida (GW/h)		
Total	Total	Total	Total	116.664	3.832.324.358	38.162.722	15.006.742	44.461.390		
América	Total	Total	Total	116.664	1.609.865.230	38.162.722	4.465.934	18.618.842		
	2006	Total	Total	14.734	23.703.254	3.553.354	0	263.452		
		Aerogeneradores	Total	Total	57	106.984	0	0	1.244	
			Sin datos	Total	Total	57	106.984	0	0	1.244
		Arietes hidráulicos	Total	Total	199	145.770	0	0	1.695	
		Hidroeléctricas	Total	Total	180	9.355.166	0	0	108.781	
		Molino de viento	Total	Total	6.512	8.828.674	0	0	102.659	
		Parque eólico	Total	Total	1	12.470	0	0	145	
		Plantas de Biogás	Total	Total	98	283.370	0	0	3.295	
		Sistemas de calentadores solares	Total	Total	1.510	3.021.610	0	0	35.135	
		Sistemas de paneles fotovoltaicos	Total	Total	6.177	902.828	0	0	10.498	
		Sin datos	Total	Total	0	1.046.382	3.553.354	0	0	
		2007	Total	Total	Total	17.231	153.036.023	3.319.843	362.869	1.767.918
		2008	Total	Total	Total	18.281	151.173.182	5.275.857	376.152	1.745.526
2009	Total	Total	Total	11.587	152.086.506	5.610.576	392.230	1.754.651		
2010	Total	Total	Total	13.007	153.677.414	4.555.676	408.580	1.775.569		

Figura 19. Interfaz de la vista de análisis Energía renovable desplegada

Vistas de análisis

La solución propuesta cuenta con una vista de análisis, con el objetivo de satisfacer las necesidades de información de la Empresa CUBAENERGIA. El Pentaho BI-Server v6.0 permite mostrar el resultado del análisis realizado, pues a través de la misma es posible ver los reportes y analizar la información mediante tablas de datos; permitiendo, desplegar el cubo de información, y modificar las vistas de análisis o crear unas totalmente nuevas. Por otra parte, esta interfaz brinda la posibilidad de imprimir el reporte o salvarlo en un archivo de formato PDF o XLS.

Reportes

En la implementación de la solución se generaron reportes candidatos útiles para la empresa CUBAENERGIA sobre energías renovables, a partir de los datos registrados.

Para la definición de los reportes se utilizó la herramienta Pentaho BI Server v6.0. Esta herramienta permite consultar los datos de distintas fuentes y ponerlos a disposición de los usuarios en diferentes formatos (HTML, PDF, Microsoft Excel, y texto plano). Las herramientas de Pentaho permiten distribuir los reportes a los usuarios interesados, así como también publicarlos para que los usuarios puedan observar la información que necesitan. A continuación, se muestra un reporte que responde a las necesidades del cliente:

Años	Países	Dispositivos generadores	Medidas
Total	Cuba	Total	194.195.236
		Aerogeneradores	7.248.682
		Arietes hidráulicos	549.110
		Hidroeléctricas	75.840.218
		Molino de viento	55.309.438
		Parque eólico	6.918.442
		Plantas de Biogás	6.986.468
		Sistemas de calentadores solares	15.497.974
		Sistemas de paneles fotovoltaicos	15.643.056
		Sin datos	10.201.848
	Artemisa	Total	391.208
	Camagüey	Total	12.391.118
	Ciego de Ávila	Total	8.746.679
	Cienfuegos	Total	49.396.247
	Granma	Total	15.558.395
	Guantánamo	Total	7.640.159
	Holguín	Total	25.250.900
	Isla de la Juventud	Total	7.597.100
	La Habana	Total	5.598.095
	Las Tunas	Total	8.549.275
	Matanzas	Total	11.690.829
	Mayabeque	Total	407.409
	Pinar del Río	Total	4.546.800
	Sancti Spiritus	Total	13.003.299
	Santiago de Cuba	Total	11.013.878
	Villa Clara	Total	10.608.761
	Ciudad de la Habana	Total	248.054

Figura 20. Obtener el ahorro de petróleo por dispositivo generador y año

3.5 Pruebas

Las pruebas se centran principalmente en la evaluación o la valoración de la calidad del producto y representan un elemento crítico para la garantía del mismo. Es una actividad en la cual un sistema o uno de sus componentes se ejecutan en circunstancias previamente especificadas, los resultados se observan, se registran y se realiza una evaluación de algún aspecto. [4]

El objetivo de la etapa de pruebas es garantizar la calidad del producto desarrollado. Además, esta etapa implica [4]:

1. Verificar que todos los requisitos se han implementado correctamente.
2. Identificar y asegurar que los defectos encontrados se han corregido antes de entregar el software al cliente.

A continuación, se exponen algunas de las pruebas que pueden ser utilizadas para la validación de un producto de software [4]:

- **Prueba unitaria:** Es el proceso de probar los componentes individuales de la solución. El propósito es identificar diferencias entre la especificación de los artefactos y el comportamiento real de cada módulo.
- **Prueba de integración:** Es el proceso en el cual los componentes son agregados para crear componentes más grandes. Es la prueba realizada para mostrar que, aunque los componentes hayan pasado satisfactoriamente las pruebas de unidad, la integración de los componentes es incorrecta.
- **Prueba de sistema:** Se refiere al comportamiento del sistema integrado. La prueba de sistema se aplica generalmente para probar los requisitos no funcionales de la solución.
- **Pruebas funcionales:** Aseguran el trabajo apropiado de los requisitos funcionales, incluyendo la navegación, entrada de datos, procesamiento y obtención de resultados.
- **Pruebas de aceptación:** El objetivo de las pruebas de aceptación es validar que un sistema cumpla con el funcionamiento esperado y permitir al usuario de dicho sistema que determine su aceptación, desde el punto de vista de su funcionalidad y rendimiento. Las pruebas de aceptación son definidas por el usuario del sistema y preparadas por el equipo de desarrollo, aunque la ejecución y aprobación final corresponden al usuario.

3.5.1 Pruebas de integración

En este tipo de pruebas se realizaron consultas a la base de datos, es necesario comprobar que el MD se poblara con los datos de los archivos fuentes y durante el proceso ETL no se haya cometido ningún

error. Los resultados que se obtuvieron fueron satisfactorios para cada una de las pruebas que se realizaron. A continuación, se muestran dos ejemplos de consultas realizadas a la base de datos.

SELECT

```
hecho_produccion_energias_renovables.capacidad_produccion,  
dimension_anno.numero_anno,  
dimension_pais.nombre_pais,  
dimension_pais.nombre_region,  
dimension_tipos_energias_renovables.nombre_tipo_fuente_energia
```

FROM

```
public.hecho_produccion_energias_renovables,  
public.dimension_pais,  
public.dimension_anno,  
public.dimension_tipos_energias_renovables
```

WHERE

```
hecho_produccion_energias_renovables.fk_dimension_pais = dimension_pais.pk_dimension_pais  
AND
```

```
hecho_produccion_energias_renovables.fk_dimension_anno = dimension_anno.pk_dimension_anno  
AND
```

```
hecho_produccion_energias_renovables.fk_tipo_energia_renovable =  
dimension_tipos_energias_renovables.pk_tipo_energia_renovable
```

ORDER BY

```
dimension_pais.nombre_region ASC,  
dimension_tipos_energias_renovables.nombre_tipo_fuente_energia ASC,  
dimension_pais.nombre_pais ASC,  
dimension_anno.numero_anno ASC;
```

Arrojando como resultado lo siguiente:

Output pane

	capacidad_produccion integer	numero_anno integer	nombre_pais character varying(50)	nombre_region character varying(50)	nombre_tipo_fuente_energia character varying(50)
921	753	2007	Angola	África	Hidroeléctricas
922	769	2008	Angola	África	Hidroeléctricas
923	769	2009	Angola	África	Hidroeléctricas
924	769	2010	Angola	África	Hidroeléctricas
925	769	2011	Angola	África	Hidroeléctricas
926	855	2012	Angola	África	Hidroeléctricas
927	855	2013	Angola	África	Hidroeléctricas
928	871	2014	Angola	África	Hidroeléctricas
929	871	2015	Angola	África	Hidroeléctricas
930	921	2016	Angola	África	Hidroeléctricas
931	276	2007	Argelia	África	Hidroeléctricas
932	276	2008	Argelia	África	Hidroeléctricas
933	276	2009	Argelia	África	Hidroeléctricas
934	276	2010	Argelia	África	Hidroeléctricas
935	276	2011	Argelia	África	Hidroeléctricas
936	276	2012	Argelia	África	Hidroeléctricas
937	276	2013	Argelia	África	Hidroeléctricas
938	276	2014	Argelia	África	Hidroeléctricas
939	276	2015	Argelia	África	Hidroeléctricas
940	276	2016	Argelia	África	Hidroeléctricas
941	1	2007	Benin	África	Hidroeléctricas
942	1	2008	Benin	África	Hidroeléctricas
943	1	2009	Benin	África	Hidroeléctricas

OK.

Figura 21. Resultado de la consulta hecha al MD

Los datos que contiene el archivo fuente con respecto a las hidroeléctricas son los siguientes:

	A	B	C	D	E	F	G	H	I	J	K
1	CAP (MW)	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
2	World	924 192	958 202	991 919	1 024 581	1 056 283	1 089 474	1 132 761	1 170 107	1 207 853	1 242 961
3	Africa	23 632	24 087	25 317	26 611	26 648	27 657	28 275	28 739	29 119	32 625
4	Algeria	276	276	276	276	276	276	276	276	276	276
5	Angola	753	769	769	769	769	855	855	871	871	921
6	Benin	1	1	1	1	1	1	1	1	1	1
7	Burkina Faso	32	32	32	32	32	32	32	32	32	32
8	Burundi	32	32	51	51	51	57	57	57	57	57
9	Cameroon	719	719	719	719	719	721	721	721	721	721
10	Cent Afr Rep	19	19	19	19	19	19	19	19	19	19
11	Comoros	1	1	1	1	1	1	1	1	1	1
12	Congo DR	2 416	2 416	2 416	2 416	2 416	2 416	2 416	2 416	2 429	2 579
13	Congo Rep	92	119	119	119	119	209	209	209	209	215
14	Cote d Ivoire	604	604	604	604	604	604	604	604	604	604
15	Egypt	2 787	2 851	2 851	2 851	2 851	2 851	2 851	2 851	2 851	2 851
16	Eq Guinea	7	7	7	7	7	127	127	127	127	127
17	Ethiopia	676	676	682	1 862	1 862	1 953	1 953	1 953	1 953	3 826
18	Gabon	170	170	170	170	170	170	330	330	330	330
19	Ghana	1 180	1 180	1 180	1 180	1 180	1 180	1 584	1 584	1 584	1 584
20	Guinea	123	123	123	125	128	128	128	128	368	368
21	Kenya	676	737	749	762	767	816	820	820	826	826

Figura 22. Archivo fuente de hidroeléctricas.

Otro de los ejemplos de las consultas realizadas a la base de datos, con respecto a la producción, el equivalente en toneladas de petróleo y los dispositivos generadores instalados por provincias es la siguiente:

SELECT

```
dimension_anno.numero_anno,  
dimension_pais.nombre_provincia,  
hecho_produccion_energias_renovables.cant_producida,  
hecho_produccion_energias_renovables.cant_dispositivos_generadores,  
dimension_dispositivos_generadores.nombre_dispositivo_generador
```

FROM

```
public.hecho_produccion_energias_renovables,  
public.dimension_pais,  
public.dimension_anno,  
public.dimension_dispositivos_generadores
```

WHERE

```
hecho_produccion_energias_renovables.fk_dimension_pais = dimension_pais.pk_dimension_pais  
AND  
hecho_produccion_energias_renovables.fk_dimension_anno = dimension_anno.pk_dimension_anno  
AND  
hecho_produccion_energias_renovables.fk_dimension_dispositivo_generador =  
dimension_dispositivos_generadores.pk_dimension_dispositivo_generador AND  
dimension_dispositivos_generadores.nombre_dispositivo_generador <> 'Sin datos' AND  
dimension_pais.nombre_provincia <> 'Sin datos'
```

ORDER BY

```
dimension_anno.numero_anno DESC,  
dimension_pais.nombre_provincia ASC,  
dimension_dispositivos_generadores.nombre_dispositivo_generador ASC;
```

CAPÍTULO 3: IMPLEMENTACIÓN Y VALIDACIÓN

Output pane

	Data Output	Explain	Messages	History	
	numero_anno integer	nombre_provincia character varying(50)	cant_producida integer	cant_dispositivos_generadores integer	nombre_dispositivo_generador character varying(50)
17	2013	Ciego de Ávila	0	1	Aerogeneradores
18	2013	Ciego de Ávila	0	0	Arietes hidráulicos
19	2013	Ciego de Ávila	3788	1	Hidroeléctrocas
20	2013	Ciego de Ávila	455	367	Molino de viento
21	2013	Ciego de Ávila	310	1	Parque eólico
22	2013	Ciego de Ávila	0	0	Plantas de Biogás
23	2013	Ciego de Ávila	73	5	Sistemas de calentadores solares
24	2013	Ciego de Ávila	59	17	Sistemas de paneles fotovoltaicos
25	2013	Cienfuegos	310	0	Aerogeneradores
26	2013	Cienfuegos	0	1	Arietes hidráulicos
27	2013	Cienfuegos	61709	16	Hidroeléctrocas
28	2013	Cienfuegos	2287	316	Molino de viento
29	2013	Cienfuegos	0	0	Parque eólico
30	2013	Cienfuegos	0	0	Plantas de Biogás
31	2013	Cienfuegos	1291	89	Sistemas de calentadores solares
32	2013	Cienfuegos	1635	120	Sistemas de paneles fotovoltaicos
33	2013	Granma	0	0	Aerogeneradores
34	2013	Granma	64	2	Arietes hidráulicos
35	2013	Granma	11832	34	Hidroeléctrocas
36	2013	Granma	2900	573	Molino de viento
37	2013	Granma	0	0	Parque eólico
38	2013	Granma	0	0	Plantas de Biogás
39	2013	Granma	1566	108	Sistemas de calentadores solares
40	2013	Granma	15207	947	Sistemas de paneles fotovoltaicos

Figura 23. Resultado de la consulta

Para el cual se tenían los archivos fuentes que a continuación se muestran:

CAPÍTULO 3: IMPLEMENTACIÓN Y VALIDACIÓN

CONCEPTO	Molino de viento	Digestores de Biogás	Plantas de Biogás	Arietes hidráulicos	Hidroeléctricas	Sistemas de calentadores solares	Sistemas de paneles fotovoltaicos	Aerogeneradores	Parque eólico
Pinar del Río	156	11	0	0	12	66	189	0	0
Artemisa	82	2	0	0	0	14	40	0	0
La Habana	138	1	0	0	0	1355	73	0	0
Mayabeque	17	0	6	0	0	57	15	0	0
Matanzas	596	161	0	0	0	11	21	0	0
Villa Clara	592	1	1	0	10	0	325	0	0
Cienfuegos	316	0	0	1	16	89	120	0	0
Sancti Spiritus	592	37	0	3	6	2	39	0	0
Ciego de Ávila	367	1	0	0	1	5	17	1	1
Camagüey	458	0		1	0	1	53	1	0
Las Tunas	966	7	1	0	0	0	368	0	0
Holguín	1051	0	33	5	13	2	595	0	1
Granma	573	0		2	34	108	947	0	0
Santiago de Cuba	29	2	3	1	25	0	822	0	0
Guantánamo	89	12	0	32	63	12	833	0	0
Isla de la Juventud	27	0	0	0	0	321	2437	11	1

Figura 24. Dispositivos instalados por provincias

CONCEPTO	Molino de viento	Digestores de Biogás	Plantas de Biogás	Arietes hidráulicos	Hidroeléctricas	Sistemas de calentadores solares	Sistemas de paneles fotovoltaicos	Aerogeneradores	Parque eólico
Pinar del Río	455,5	8,8	0	0	130,3	95,7	393,0	0	0
Artemisa	56,5	15,5	0	0	6,3	20,3	63,2	0	0
La Habana	95,1	22,5	0	0	0	1964,8	60,6	0	0
Mayabeque	11,7	0	0	0	0	82,7	3,2	0	0
Matanzas	861,7	861,7		1,0	0	16,0	23,4	0	0
Villa Clara	843,1	0	593,0	8,0	77,4	0	546,7	0	0
Cienfuegos	228,7	0	0	0	6 170,9	129,1	163,5	31,0	0
Sancti Spiritus	1146,4	7,1	0	0	762,0	2,9	31,1	0,3	0
Ciego de Ávila	45,5	112,0	0	0	378,8	7,3	5,3	0	31,0
Camagüey	1299,0	0	0	1,0	0	1,5	125,5	0	0
Las Tunas	721,8	0	0	0,4	0	0	599,9	0	0
Holguín	1601,7	0	1822,0	0,2	972,0	2,9	1231,6	0	1 786,5
Granma	290,0	0		6,4	1183,2	156,6	1520,7	0	0
Santiago de Cuba	28,1	0	1,8	0	800,3	0	1172,1	1808,0	0
Guantánamo	36,0	0	0	0	463,6	17,4	1873,3	0	0
Isla de la Juventud	10,9	0	0	0	0	465,5	1852,7	0	119,7

Figura 25. Energía producida por dispositivos instalados.

3.5.2 Matriz de trazabilidad

Un concepto clave en el proceso de gestión de cambios es la trazabilidad. Los requisitos deben ser trazables, es decir, “rastreables”. Se podría decir que un requisito es trazable si se pueden identificar todas las partes del producto existente relacionadas con ese requisito. Todos los requisitos deberían ser trazables para mantener consistencia entre los distintos documentos de un proyecto. [29]

Una de las claves para la gestión de cambio es la matriz de trazabilidad, la cual permite establecer las relaciones entre los diferentes requisitos desde un punto de vista jerárquico y la documentación generada durante todo este proceso es la base sobre la que debe construirse todo el plan de pruebas. El uso de matrices de trazabilidad es una buena técnica para llevar a cabo esta actividad de forma eficiente. [29]

Para este proyecto se define la matriz de trazabilidad para los requisitos de información con respecto al hecho y a las dimensiones definidas. Para lograr hacer un seguimiento de los requisitos de información y tener una prueba de que las dimensiones se relacionan perfectamente con los requisitos definidos, se realizó la siguiente matriz de trazabilidad.

(12) Requisitos de información		RI 01	RI 02	RI 03	RI 04	RI 05	RI 06	RI 07	RI 08	RI 09	RI 10	RI 11	RI 12
By: <input type="text" value="Transitor"/>													
(5) Dimensiones													
<input type="checkbox"/> dimension_anno		✓	✓	✓		✓		✓	✓		✓		
<input type="checkbox"/> dimension_dispositivos_generadores		✓	✓			✓							
<input type="checkbox"/> dimension_pais		✓			✓		✓		✓	✓		✓	
<input type="checkbox"/> dimension_tipos_energias_renovables				✓	✓			✓	✓	✓	✓		✓
<input type="checkbox"/> hecho_produccion_energias_renovables		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Figura 26. Matriz de trazabilidad

3.6 Conclusiones del capítulo

A partir del diseño propuesto se implementó el Mercado de Datos de Energías Renovables. El desarrollo de los subsistemas que componen la solución permite realizar el análisis de las energías renovables tanto a nivel nacional como internacional y la generación de reportes. Se realizaron pruebas para comprobar que los datos que se encontraban en las fuentes de información fueron todos cargados en el MD. Además se realizó una matriz de trazabilidad para representar la relación existente entre los requisitos de información y las dimensiones.

Conclusiones generales

La investigación realizada cumple los objetivos planteados inicialmente mediante el desarrollo del Mercado de datos Energías Renovables, arribando a las siguientes conclusiones:

- ✓ Se evidenció la necesidad de desarrollar un MD para el análisis de los datos publicados en diferentes fuentes sobre energías renovables para el apoyo a la toma de decisiones utilizando la metodología propuesta por DATEC, el tipo de almacenamiento ROLAP y el esquema estrella.
- ✓ Se obtuvo el modelo de datos, el diseño del cubo OLAP y de las transformaciones a partir de los requisitos de información definidos.
- ✓ La solución implementó los subsistemas de integración, almacenamiento y visualización permitiendo el análisis sobre las energías renovables y la generación de reportes.
- ✓ A partir de las pruebas realizadas se comprobó que el MD se pobló con los datos de los archivos fuentes y que durante el proceso ETL no se cometió ningún error. También se realizó la matriz de trazabilidad para la representación de los requisitos de información con cada una de las tablas de dimensiones y con la tabla de hechos.

Recomendaciones

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

- ✓ Trazar una estrategia de carga incremental de los datos, en caso de actualizar el MD con nuevos datos que se publiquen.

Referencias Bibliográficas

1. Sedano, C.P.R., Repositorio Digital de Tesis PUCP., 2009.
2. Fernando. *Energías Renovables*. 2015; Available from: <http://erenovable.com/energias-renovables/>.
3. Bernabeu, R.D., *Data Prix*. 2010.
4. Hernández, I.Y., *PROPUESTA DE METODOLOGIA PARA EL DASARROLLO DE AD EN DATEC*. 2010.
5. Asenjo, J.S., *Sistema Gestores de Base de Datos*. . 2009.
6. Llosas, Y.M.P., *Herramienta informática para la gestión de la calidad del posgrado*. REVISTA RIEMAT, 2016. **1**.
7. Rivadera, G.R., *La metodología de Kimball para el diseño de almacenes de datos*. . 2010.
8. González, Y.d.I.Á.G.P.y.M.T.R., *Mercados de datos para el análisis estadístico de la información*. 3C tecnología, 2014. **3**.
9. KONCILIA, R.W.Y.C., *DATA WAREHOUSES AND OLAP*. 2007.
10. García, J.H.M., *La inteligencia de negocios como herramienta para la toma de decisiones estratégicas en las empresas. Análisis de su aplicabilidad en el contexto corporativo colombiano*. 2010.
11. Ralph Kimball, M.R., Warren Thornthwaite, Joy Mundy, Bob Becker, *The Data Warehouse Lifecycle Toolkit*. 2002.
12. SINNEXUS, *Datamart*.
13. Kimball, R., *The Data Warehouse Lifecycle Toolkit*. 2008.
14. Curto, J. *Data Warehousing, Data Warehouse y Datamart*. 2007.
15. E.F. Codd, S.B.C.a.C.T.S., *Providing OLAP to user-analysts: An IT mandate*. 1993.
16. CASTILLO, D.O.A.F.Y.J.N.P., *Estado actual de las tecnologías data warehousing y OLAP aplicadas a bases de datos espaciales* Luís Joyanes Aguilar.
17. Molier, F.J.L., *Grupo A y B Informatica :Bloque específico*. 2005. **Vol. II.**: p. 173,174,175.
18. Tamayo, M.y.M., Francisco Javier., *Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP*. Sistema de Información Científica, 2006. **Vol. 26**.
19. VÁZQUEZ, F.P.Y.G., *Relevamiento: Diseño Físico de Sistemas OLAP*.
20. *Esquema Físico Molap-Rolap-Holap*. Available from: http://www.interaktiv.cl/blog/wp-content/uploads/2012/04/8-Esquema_FisicoMolap-Rolap-Holap.pdf.
21. Borrell Madrazo, T., Gómez Romero, Yusdaily, Barbán Góngora, Nilber, *Análisis, diseño e implementación del mercado de datos del indicador medio ambiente para el almacén de datos de la Oficina Nacional de Estadísticas*. 2010, UCI.
22. Rivadera, G.R. *La metodología de Kimball para el diseño de almacenes de datos (Data warehouses)* 2010.
23. González, Y.P., *Sistema de información de gobierno Mercado de datos Inmigración y extranjería*. 2011.
24. Dataprix., *OLAP, MOLAP y ROLAP*.
25. Sommerville, I., *Ingeniería del software*. 2005.
26. Team, V.P. *Il-in-One, End-to-End Information Technology System Modeling Tool*. 2014; Available from: <http://www.visual-paradigm.com/product/vpuml/editions/enterprise.jsp>
27. Corporation., P.P., *Powerful Analytics Made Easy*. 2014.
28. Fonseca, K.F.F.y.C.N., *Análisis, diseño e implementación del mercado de datos para los indicadores sobre el Plan Turquino para la Oficina Nacional de Estadísticas*. 2010, Universidad de las Ciencias Informáticas.
29. Lina María Montoya-Suárez, J.C.M.-G., Jorge Mauricio Sepúlveda-Castaño, *Un Caso de Estudio para la Adopción de un Modelo de Trazabilidad de Requisitos en el Sector Energético*. 2014.

Anexos

Anexo # 1: Vista principal de la aplicación.



Figura 27. Vista principal de la aplicación

Anexo # 2: Gráfico de análisis regional.

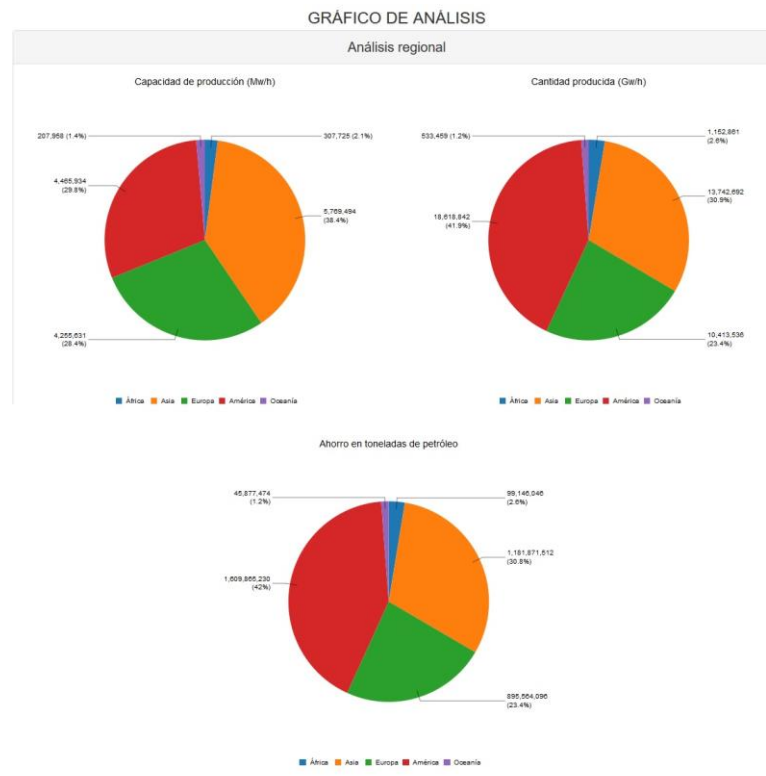


Figura 28. Gráfico de análisis regional

Anexo # 3: Gráfica de análisis sobre el comportamiento internacional de la cantidad de energía producida anualmente.

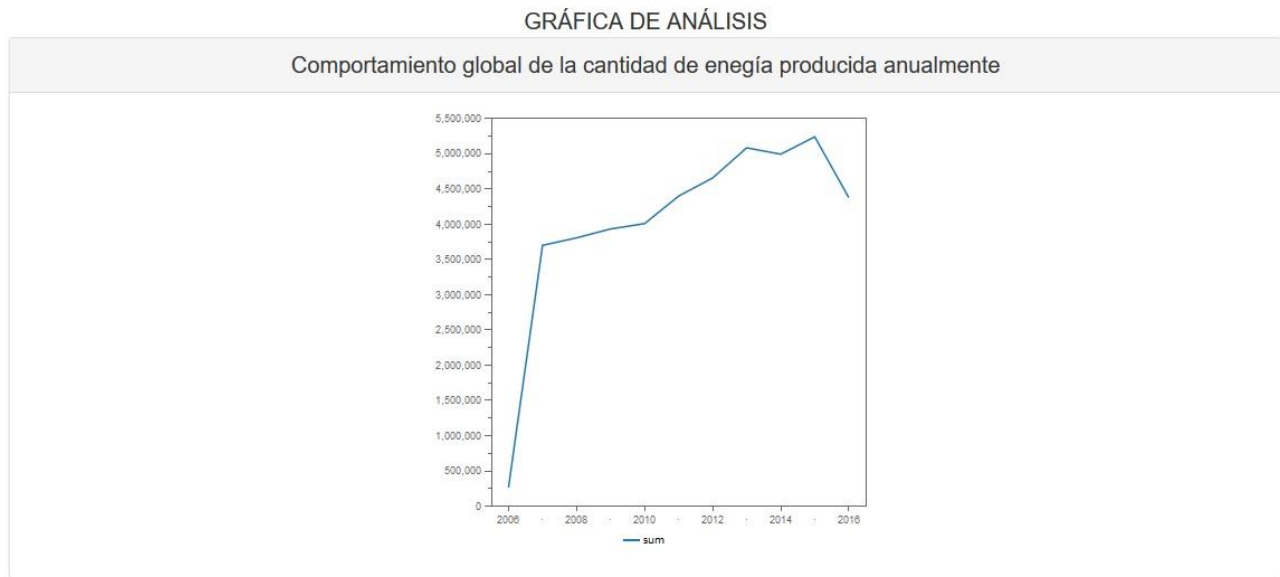


Figura 29. Comportamiento internacional de la cantidad de energía producida anualmente.

Anexo # 4: Gráfica de análisis de cantidad de energía producida anualmente filtrado por región, país y tipo de energía renovable.

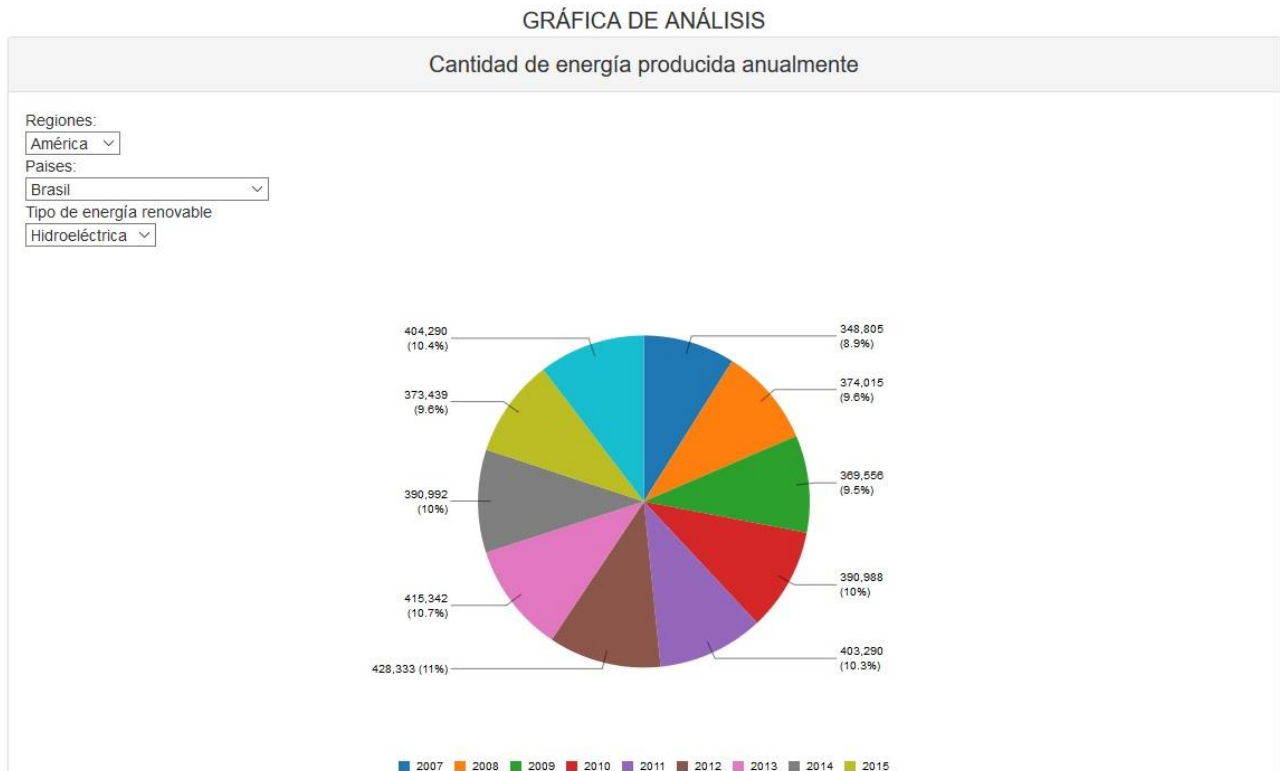


Figura 30. Gráfica de análisis de cantidad de energía producida anualmente

Anexo # 5: Gráfica de análisis de ahorro en petróleo por provincia y años.

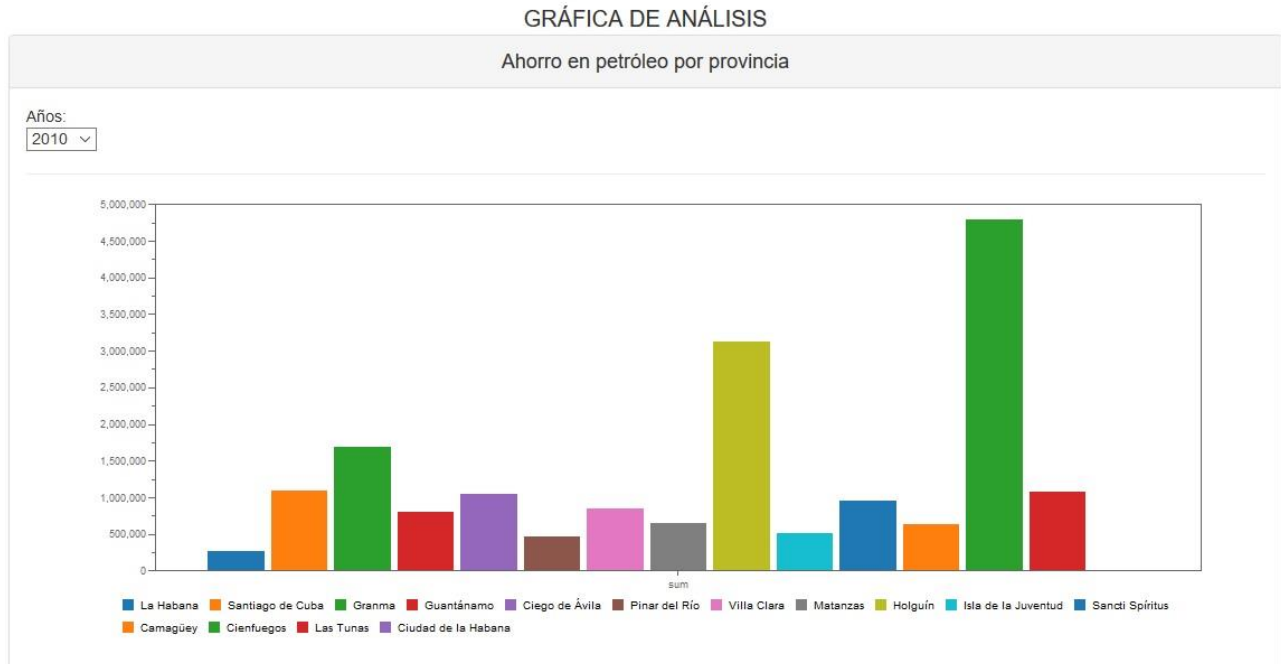


Figura 31. Gráfica de análisis de ahorro en petróleo

Anexo # 6: Gráfica de análisis de ahorro en petróleo seleccionando una provincia.

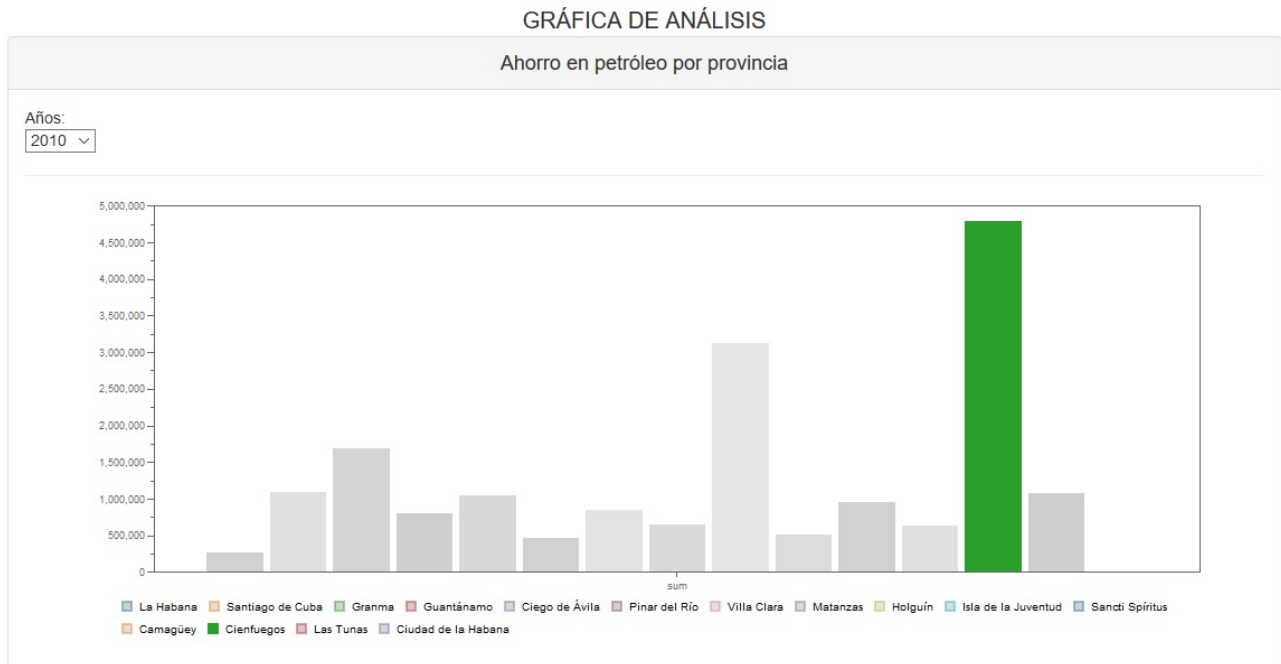


Figura 32. Gráfica de análisis de ahorro en petróleo seleccionando una provincia

Anexo # 7: Gráfica de consumo físico de la provincia seleccionada en el anexo 6.

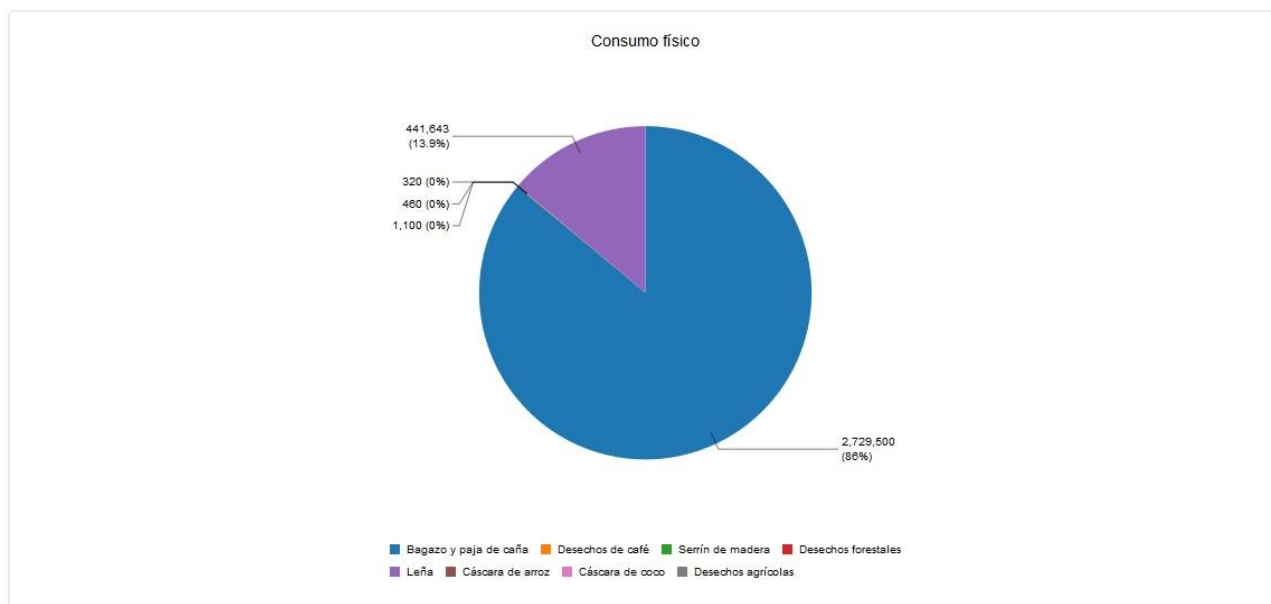


Figura 33. Gráfica de consumo físico de la provincia

Glosario de términos

Aerogeneradores: Son molinos de viento con aspas o paletas de gran diámetro que se mueven con la acción del aire, y a su vez accionan un generador o dinamo el cual produce energía eléctrica.

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

Arietes hidráulicos: También llamado Carnero, es un sistema mecánico que utiliza como fuente de energía una corriente de agua y que permite bombear agua a determinadas alturas.

Atributo: Es cada una de las cualidades, propiedades o características de un elemento.

Base de datos históricos: Se exponen los principales componentes de las bases de datos como genuinos ejemplos de los sistemas de información histórica y las fuentes de información histórica.

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

BI: Inteligencia del negocio.

Biomasa: Se considera incluida en esta categoría toda materia vegetal creada por la fotosíntesis y toda materia orgánica contenida en los residuos animales, industriales o domésticos.

Calentadores Solares: Todo sistema que, empleando la energía del sol, permita calentar agua a temperaturas superiores a la ambiental

CUBAENERGIA o Centro de Gestión de la Información y Desarrollo de la Energía: tiene como misión gestionar información en energía, aplicaciones nucleares y tecnologías de avanzada; desarrollar productos y tecnologías para la sostenibilidad energética y realizar acciones de divulgación que contribuyan a potenciar la cultura científica en el país.

DATEC: Centro de Tecnologías de Gestión de Datos.

Digestores de biogás: Cualquier sistema capaz de producir gas combustible mediante la fermentación anaerobia de biomasa, sin el empleo de dispositivos mecánicos.

Energía Eólica: Es la energía que se genera por diferencia de temperatura en la atmósfera.

Energía Primaria: Se designan así a las fuentes que únicamente implican proceso de extracción o de captación, con o sin separación del material contiguo. Es provista por la naturaleza de forma directa como la Energía Solar, Eólica e Hidráulica; después de someterse a un proceso de extracción como el petróleo, el gas natural y el carbón mineral o a través de los procesos de fotosíntesis, como la biomasa.

Energía Solar: Es la obtenida de la radiación solar. Para aprovechar este tipo de energía, se dispone de dos vías: La conversión Térmica y la conversión Fotovoltaica. Con estos fines se usan los Sistemas de Calentadores Solares, Sistemas de Paneles fotovoltaicos, Destiladores Solares, etc.

ETL: proceso de extracción, transformación y carga.

Fuentes de Energía Renovables: Son fuentes de energía no convencionales, que sustituyen a los portadores tradicionalmente usados, petróleo y derivados. Estas Fuentes se caracterizan por la capacidad de renovarse ya sea permanente o de forma periódica, entre ellas se encuentran:

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

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.

Hidroeléctricas: Sistemas accionados por la fuerza del agua y capaces de generar electricidad.

Hidroenergía: Es la energía potencial que se deriva de la caída del agua.

HOLAP: Procesamiento Analítico Híbrido en Línea.

IRENA o Agencia Internacional de Energías Renovables: es una organización intergubernamental que apoya a los países en su transición hacia un futuro energético sostenible y sirve como la plataforma principal para la cooperación internacional, un centro de excelencia y un repositorio de políticas, tecnología, recursos y conocimiento financiero sobre energías renovables.

Malacates: Son bombas mecánicas que emplean como energía la fuerza animal para el accionamiento de pistones, que son los encargados de mantener un caudal de agua a través de una tubería. Se emplean fundamentalmente en el bombeo de agua.

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

MOLAP: Procesamiento Analítico Multidimensional en Línea.

Molino de viento: Dispositivo que emplea la energía de aire (eólica) y la transforma en energía mecánica.

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

ONEI o La Oficina Nacional de Estadística e Información: creada por el artículo 31 del Decreto Ley No. 281 del 2 de febrero de 2011, como resultado de la organización del Sistema de Información del Gobierno presenta entre sus informes, estadísticas sobre energías renovables.

Open Source (código abierto): Práctica de desarrollo de software que promueve el acceso al código fuente de los sistemas computacionales.

Paneles fotovoltaicos: Sistema que transforma la energía de la radiación solar en energía eléctrica, mediante paneles captadores de los rayos solares que contienen en su superficie elementos semiconductores, que realizan esta transformación.

Parque eólico: Son extensas áreas con generadores de turbinas de viento de tamaño mediano (de 15 a 30 metros de diámetro, con una potencia entre 100 y 400 kW) instalados en filas para la producción de energía a gran escala con un rendimiento satisfactorio. En los mayores parques eólicos del mundo sus turbinas pueden generar unos 1 120 MW de potencia (una central nuclear puede generar unos 1 100 MW).

Plantas de biogás: Cualquier sistema capaz de producir gas combustible mediante la fermentación anaerobia de biomasa, con el empleo de dispositivos mecánicos

ROLAP: Procesamiento Analítico Relacional en Línea.

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.

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.