

Facultad 8



Título: Desarrollo de un Almacén de Datos para el Control del Consumo de Portadores Energéticos en la Oficina Nacional de Estadística.

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

Autor

René Raydel Escobar Domínguez.

Tutores

Ing. Mabel Medina Rodríguez.

Ing. Yanisbel González Hernández.

Ciudad de La Habana, junio 2010.



Tenemos que cambiar la tradicional actitud ante la construcción de software. En vez de pensar que nuestra principal tarea es indicar a un ordenador qué hacer, concentrémonos en explicar a las personas lo que queremos que el ordenador haga...

-- Donald E. Knuth --

DECLARACIÓ DE AUTORÍA

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

Para que así conste firmo la presente a los 22 días del mes de Junio del año 2010.

Firma del Autor

René Raydel Escobar Domínguez.

Firma del Tutor

Ing. Mabel Medina Rodríguez.

Firma del Tutor

Ing. Yanisbel González Hernández.

Ing. Mabel Medina Rodríguez.

Categoría Docente: Instructor

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

Dos años como profesora del Departamento de Sistemas Digitales de la Facultad 4 y 1 año como especialista del grupo de Arquitectura de Información de la Dirección Técnica de la Infraestructura Productiva.

Ha impartido las asignaturas de Teleinformática.

Actualmente trabaja en el Centro de Tecnologías de Gestión de Datos de la Infraestructura Productiva.

Fundadora de la Universidad de Ciencias Informáticas.

Ha tutorado dos tesis de pregrado con resultados satisfactorios.

Se encuentra matriculada en la maestría de Gestión de Proyectos Informáticos, donde ha cursado más del 80% de los cursos.

Ing. Yanisbel González Hernández

Categoría Docente: Instructor Recién Graduado

Graduado en el 2008 de Ingeniero en Ciencias Informáticas de la Universidad de Ciencias Informáticas (UCI).

Un año como profesora del Departamento de Ciencias Sociales y Humanísticas de la Facultad 7 y 4 meses como especialista de la Dirección de Producción 1 de la Infraestructura Productiva (IP) y después se pasa a ser Analista de la Línea de Soluciones de Almacenes de Datos e Inteligencia de Negocio del Centro de Tecnologías de Gestión de Datos de la (IP).

Fundadora de la Universidad de Ciencias Informáticas.

Ha tutorado una tesis de pregrado con resultados satisfactorios.

Se encuentra matriculada en la maestría de Gestión de Proyectos Informáticos, donde ha cursado más del 80% de los cursos.

A mis padres René y Sondra, por ser mis ejemplos, mis amigos incondicionales, por estar en el momento en que más los necesito y brindarme su apoyo en difíciles decisiones que en este largo andar de la vida me ha impuesto el destino.

A mi abuelo Pablo, ejemplo que siempre he seguido, figura que ha marcado en mí, el respeto a la vida y al ser humano. A mi desaparecida abuela que con tanto amor junto con mi abuelo contribuyeron a mi educación. Este logro es parte del fruto que cosecharon.

A mis hermanas, que ambas son unos de los tesoros más preciados para mí, la más chica Rena... espero que crezca con sabiduría, a Yulia, ya que casi es una profesional, que siga con su buen avance de ser mejor cada día, espero lo mejor de ellas.

A mi padrastro, a su familia, en especial a José. Al resto de mi familia, a Vina mi abuela por amor y a Zamora mi querido padrino.

A mis amistades de la UCI que han sido una familia para mí desde que entre a la universidad, hemos convivido por 5 años, si los menciono no cabrían en este documento, pero siempre hay algunos en especial, Marislav, Alianne, Leydis, Kenia, Yanitza, Dania, Leticia, Raul, Yuri, Jáilen, Roberto, Abel en fin todos aquello que de alguna forma han compartido conmigo buenos y malos momentos. Gracias y éxitos para todos.

A mis tutoras, Yanisbel y Mabel, que han sido capaz de ir moldeando mi forma de proyectarme y de ser mejor cada día, siempre han creído en mí y que puedo dar más de lo que doy, gracias en verdad sin ustedes no habría podido lograr este fruto.

A mi tribunal que con sus revisiones han logrado que esta investigación sea digna de un profesional por el que estoy luchado ser, especialmente Asnioby por su incalculable ayuda en la investigación, no sin dejar atrás su gran forma de ser, que es capaz de brindarte una mano después de una caída.

En fin a todas aquellas personas que de alguna que de otra forma han contribuido con el éxito de esta investigación.

Dedico esta investigación a todas aquellas personas que hayan convivido en algún momento el mismo espacio y tiempo que yo, principalmente a mi familia siendo el motor inspirador de mis logros como persona y seguimientos como estudiante. Más que eso, aquellas que han sido mis grandes modelos a seguir, mi papá, mi padrastro, mi abuelo, a la memoria de mi abuela y fundamentalmente a la persona que dio paso a mi existencia en este mundo, mi mamá. Siendo ella junto con mis hermanas Rena y Yulia los seres más preciados e importantes para mí en la vida.

Les dedico este logro a ustedes...

Resumen

El presente trabajo de diploma se enmarca en el área de los almacenes de datos para el análisis de información relacionada con los Portadores Energéticos. Para guiar el desarrollo de la investigación se toma la metodología que se utiliza en el Centro Tecnología de Gestión de Datos (DATEC) para este tipo de soluciones.

Como resultados se exponen las estructuras dimensionales para el modelo propuesto de Portadores Energéticos de la Oficina Nacional de Estadísticas (ONE) que comprende las dimensiones, jerarquías, tablas de hechos y medidas necesarias para soportar los análisis de consumo de energía eléctrica y de combustible. Igualmente se definen e implementan los mecanismos de extracción, transformación y carga de los datos correspondientes al modelo propuesto, así como se desarrolla la capa de Inteligencia del Negocio (BI) para la visualización y la realización de análisis de la información contenida en el Almacén de Datos para el Control del Consumo de Portadores Energéticos. Para lograr la aceptación de cliente y una buena calidad del producto se trazan además un conjunto de pruebas.

Su principal novedad consiste en su implementación utilizando herramientas libres.

Palabras claves: almacenes de datos, estructuras dimensionales, jerarquías, ONE, BI, Portadores Energéticos.

Índice

Resumen	VI
Introducción	2
Capítulo I: Fundamentación Teórica	8
Introducción	8
1.1 Portadores Energéticos	8
1.1.1 Tipos de Portadores Energéticos más Comunes son:	8
1.2 Indicadores	9
1.2.1 Principales Indicadores Energéticos En el Mundo	9
1.2.2 Principales Indicadores Energéticos En Cuba.....	9
1.3 Herramientas Existentes para el Control de Energía.....	10
1.3.1 Algunas de las Herramientas Existentes En el Mundo	10
1.3.2 Algunas de las Herramientas Existentes En Cuba	10
1.3.3 Análisis de las Soluciones Existentes.....	11
1.4 Tecnologías de Almacenamiento de Datos	11
1.4.1 Base de Datos.....	11
1.4.2 Almacenes de Datos (<i>DataWareHouse</i>).....	12
1.4.3 Mercados de Datos (<i>DataMart</i>).....	13
1.4.4 Justificación de la Tecnología de Almacenamiento a Utilizar	14
1.5 Metodologías para el desarrollo	15
1.5.1 Justificación de la Metodología a Utilizar	16
1.6 Modelos de Datos	17
1.6.1 Modelo Entidad-Relación.....	17
1.6.2 Modelo Dimensional.....	18
1.7 Herramientas de Modelado	20
1.7.1 Justificación de la Herramienta CASE a Utilizar	21
1.8 Gestores de Base de Datos.....	22
1.8.1 Justificación del Gestor de Base de Datos a Utilizar	22

1.9 Modos de Almacenamiento de Datos	24
1.9.1 Procesamiento Analítico Relacional (ROLAP).....	24
1.9.2 Procesamiento Analítico Multidimensional (MOLAP).....	24
1.9.3 Procesamiento Analítico Híbrido (HOLAP).....	25
1.9.5 Justificación del Modo de Almacenamiento Seleccionado	25
1.10 Herramientas para el proceso ETL.....	25
1.10.1 Justificación de la Herramienta de Integración a Utilizar	27
1.11 Herramientas para el proceso BI	27
1.11.1 Justificación de las Herramientas para los Procesos de BI.....	28
1.12 Herramientas seleccionadas para la realización de pruebas	29
1.13 Conclusiones del Capítulo	31
Capítulo II: Análisis y Diseño	32
Introducción	32
2.1 Análisis de la Solución	32
2.1.1 Definición del Negocio	32
2.1.2 Tema de Análisis Identificado	33
2.1.3 Roles y Permisos	33
2.1.4 Reglas del Negocio	33
2.1.5 Necesidades de los Usuarios.....	34
2.1.6 Requisitos de Información	34
2.1.7 Requisitos Multidimensionales	35
2.1.8 Requisitos Funcionales.....	35
2.1.9 Requisitos no Funcionales.....	36
2.1.10 Casos de uso del sistema	36
2.2 Diseño de la Solución	36
2.2.1 Matriz BUS o Matriz Dimensional	36
2.2.2 Modelo de Datos.....	37
2.2.3 Arquitectura de los Componentes del Sistema.....	38

2.4 Conclusiones del Capítulo	40
Capítulo III: Implementación y Pruebas.....	41
Introducción	41
3.1 Modelos de Datos Físico.....	41
3.1.1 Estructuras de Datos	41
3.1.2 Procesos Extracción Transformación y Carga (ETL).....	45
3.1.3 Usuarios y Privilegios	47
3.2 Guía de Implantación	48
3.2.1 Requerimientos.....	49
3.2.2 Secuencia de Pasos	49
3.4 Procesos de Inteligencia del Negocio (BI)	50
3.5 Validación y pruebas	52
3.5.1 Listas de Chequeo Análisis	52
3.5.2 Listas de Chequeo de Diseño	53
3.5.3 Validación de requisitos por el cliente	53
3.5.4 Pruebas de Rendimiento y Estrés	53
3.5.5 Pruebas de Volumen y Carga	56
3.5.6 Pruebas de Implantación	58
Conclusiones del Capítulo	59
Conclusiones	60
Recomendaciones	60
Bibliografía.....	61
Glosario de Términos	65
Anexos.....	66

Introducción

La humanidad ha alcanzado un considerable desarrollo tecnológico a gran escala siendo la segunda mitad del siglo XX el período de más crecimiento en muchos sectores científicos. Es indiscutible que la informática ha tomado el control en cuanto a adelanto técnico se refiere, convirtiéndose en una ciencia clave para el progreso de la humanidad y de excepcional apoyo para otras ramas tecnológicas.

Gracias a estos avances, los sectores del Software y servicios informáticos que se engloban dentro de lo que se conoce como la Industria de las Tecnologías de la Información, particularmente en los últimos años, se ha convertido en un elemento esencial para el aumento de la economía y otros renglones en varios países, siendo los software estadísticos una fuerte base para apoyar estas esferas. En la actualidad existe una estrecha unión de la Estadística con la Informática. Esta vinculación tiene un gran significado puesto que la estadística entre otras cosas permite determinar problemas antes de que estos ocurran, evaluar nuevos procedimientos y su impacto en la producción surgiendo la necesidad de implementarlos de alguna manera en forma de herramientas informáticas, o de realizar estudios de simulación para determinar algunas propiedades del mismo.

En Cuba el desarrollo de la Industria del Software es un medio importante para incrementar el desempeño económico, pues requiere menos inversión que las restantes, tomando en consideración que la inversión principal ya se ha realizado a partir de la política educacional que ha seguido la Revolución desde 1959.

Sin embargo, en la actualidad aún existen un gran número de instituciones que se encuentran envueltas en un gran flujo de información estadísticas, lo que provoca cierta lentitud en el tratamiento de dicha información. El uso del Software en la Estadística es muy importante a la hora de resolver este tipo procesamiento de información; ya que en la mayoría de los casos facilita los cálculos matemáticos y la elaboración de gráficos.

Una de las instituciones que ha ido mitigando este tipo de procesamiento de información apoyándose en estos tipos de software estadísticos, es La Oficina Nacional de Estadística, órgano rector de la estadística en Cuba, cuyo principal objetivo se fundamenta en captar, analizar y difundir los datos recogidos a lo largo y ancho de todo el País. Para ello cuenta con una serie de modelos estadísticos en los cuales recoge información de todos los sectores de la economía y la sociedad.

Unos de estos sectores es el control del consumo de portadores energéticos el cual se caracteriza por ser

un sistema basado en estadísticas continuas sobre producción y consumo. Este sector es la razón fundamental por la que actualmente en el país se lleva a cabo La Revolución Energética como una necesidad y alternativa de ahorro del país. Dentro de las transformaciones de esta se puede citar la reducción del consumo eléctrico de la vivienda y con ello la mejora de la economía familiar, la prolongación de las reservas de combustibles fósiles del país y la disminución del impacto ambiental provocado por el uso irracional.

Uno de los principales problemas que presenta la Oficina Nacional de Estadísticas es la carencia de información relacionadas con el Control del Consumo de Portadores Energéticos de las distintas oficinas esparcidas por todo el País. Los datos enviados por estas oficinas regionales en ocasiones sólo son un resumen de los indicadores reales que se deberían medir, además el formato de estos documentos es en Excel, provocando un engorroso trabajo de forma manual a la hora de realizar algún tipo de análisis, porque se interactúa con una gran cantidad de documentos, aumentando las probabilidades de errores estadísticos. La no existencia de una forma adecuada de almacenar, recuperar y presentar la información proveniente de estos organismos regionales, dificulta el análisis de los principales reportes y cruces de variables, indicadores, porcentajes y demás aspectos de interés; obstaculizando la disponibilidad de información sobre portadores energéticos para Órganos del Estado y afectando el proceso de toma de decisiones.

Partiendo de lo anteriormente expuesto, se plantea el siguiente **Problema de Investigación** ¿Cómo mejorar el almacenamiento de datos, para optimizar el análisis de la información almacenada en la Oficina Nacional de Estadísticas sobre el Control del Consumo de Portadores Energético de las distintas oficinas regionales?

Determinando como **Objetivo General**: Desarrollar la estructuración de los planes y el Control del Consumo de Portadores Energéticos en un modelo dimensional, para flexibilizar el análisis y el proceso de toma de decisiones en la Oficina Nacional de Estadísticas.

Por tanto el **Objeto de Estudio** de este trabajo son los Almacenes de Datos, estando la investigación enmarcada en el **Campo de Acción** de los Almacenes de Datos de temas de energéticos.

Para dar cumplimiento al objetivo general se plantean los siguientes **Objetivos Específicos**:

1. Elaborar el marco teórico de la Investigación acerca de las principales tendencias de implementación de los almacenes y estudio del arte del Control del Consumo de Portadores

Energéticos.

2. Analizar el Modelo para Control del Consumo de Portadores Energéticos en la Oficina Nacional de Estadística.
3. Diseñar el Almacén de Datos para el Control del Consumo de Portadores Energéticos en la Oficina Nacional de Estadística.
4. Cargar los datos en el Almacén de Datos sobre el Control del Consumo de Portadores Energéticos para la Oficina Nacional de Estadística.
5. Implementar la presentación visual de la información.
6. Realizar pruebas de funcionalidad para validar la solución desarrollada.

Surgiendo como **Idea a defender**:

Si se logra la estructuración de los planes y el Control de Consumo de Portadores Energéticos en un modelo dimensional, para flexibilizar el análisis y el proceso de toma de decisiones en la Oficina Nacional de Estadísticas, se podrá agilizar el proceso de realización de los principales reportes y cruces de variables, indicadores, porcentajes y demás aspectos de interés.

Para dar seguimiento a los objetivos específicos, estos fueron desglosan en las siguientes **Tareas de Investigación**:

1. Documentar los temas relacionados con el desarrollo de Almacenes de Datos para desarrollar una solución con mayor calidad que cumpla con todos los requerimientos necesarios.
2. Caracterizar la metodología a utilizar en el desarrollo lo cual permite asegurar que dicha metodología es la más adecuada para dar respuesta a las necesidades de la solución.
3. Identificar estructura de usuarios y permisos para que la solución cuente con una buena seguridad permitiendo que solo accedan los usuarios pertinentes y que los mismos posean los privilegios adecuados.
4. Definir temas de análisis lo que permite enfocarse en las áreas específicas que abarca la solución.
5. Identificar necesidades de información, requisitos funcionales y no funcionales para detallar cuáles son aquellos elementos funcionales que debe poseer la solución.
6. Modelar requerimientos con el fin de que dichos requerimientos se adapten mejor a las necesidades de la solución.
7. Validar requerimientos, para que la solución cuente con las exigencias que realmente son

necesarias para su funcionamiento.

8. Definir requisitos de entrada y de salida para tener conocimiento de las variables tanto de entrada como de salida que conformarán la solución.
9. Elegir la granularidad del proceso del negocio.
10. Definir las dimensiones del Almacén de Datos, las cuales contienen la descripción textual del negocio lo que permite poseer los elementos necesarios para elaborar el diseño lógico y físico del almacén de datos.
11. Definir los hechos asociados a las dimensiones definidas, donde se almacena la información concerniente al tema en cuestión, es decir las dimensiones.
12. Estructurar el modelo dimensional lo que permite definir una manera de organizar las dimensiones en un modelo.
13. Transformar del modelo dimensional al diseño físico con el fin de facilitar la implementación de la solución.
14. Implementar la Base de Datos para poder realizar todas las operaciones que contiene la solución.
15. Implementar los cubos OLAP con el fin de organizar la información dentro de la base de datos.
16. Definir estrategia de Extracción, Transformación y Carga de los datos.
17. Montaje de los datos para el Control del Consumo Energético en el almacén de datos de la Oficina Nacional de Estadística.
18. Realizar los listados de chequeos para controlar el desarrollo de la aplicación.
19. Realizar pruebas de volumen y estrés al Almacén de Datos y a la funcionalidad a fin de comprobar la eficiencia de ambos.

En la realización de este trabajo se utilizan los siguientes **Métodos Científicos**:

- **Métodos Teóricos:**

1. **Analítico-Sintético:** Posibilitó sintetizar las características esenciales, permitió desglosar las características y funcionalidades de las técnicas de modelación de procesos de negocio, para identificar los beneficios que ofrecen y seleccionar una que sea lo más completa posible, que permita capturar los requisitos eficientemente para el posterior análisis del Control del Consumo Energético en la Oficina Nacional de Estadísticas, como también permitió arribar a las conclusiones de la investigación.
2. **Histórico-Lógico:** Permitió estudiar la evolución de los Modelos Estadísticos para el Control del Consumo Energético y las diferentes formas de Almacenamiento de Datos. Así como hizo posible un estudio histórico sobre el proceso de captura de requisitos a través de diversas técnicas y metodologías de desarrollo de software, permitió comprender la necesidad de aplicar alguna o todas sus combinaciones y cómo se hace, permitiendo tener un mayor conocimiento para aplicarlo en la solución del problema presentado.
3. **Modelación:** Este método permitió conformar el proceso del negocio y predecir la respuesta de dicho proceso a variaciones de algunos de sus parámetros sin tener que ejecutarlo en la realidad.

- **Métodos Empíricos:**

1. **Observación:** Se utilizó para percibir los problemas existentes con relación al Control del Consumo Energético, ayudó a obtener una mejor visión del problema y el objetivo de esta investigación.
2. **Entrevista:** Es usado para conocer las funcionalidades que debe tener el sistema a desarrollar teniendo en cuenta las necesidades del cliente. Se usa al inicio y durante todo el desarrollo del trabajo, pues a medida que se desarrolla el sistema se analiza si el mismo cumple las expectativas requeridas.

Estructura de la Tesis

El presente trabajo está compuesto por 3 capítulos, de los cuales el primero aborda los temas relacionados con la fundamentación teórica, el segundo trata sobre la descripción de la solución y el tercero está orientado al análisis y validación de los resultados.

El Capítulo 1, referido al marco teórico y referencial de la investigación, así como técnicas, metodologías y software usados para apoyar la solución del problema.

En el Capítulo 2 se abordan aspectos concernientes a la descripción de la solución, específicamente, a la descripción de las fuentes a integrar, definición de las áreas de análisis, la arquitectura, el diseño, la interacción entre los componentes, la implementación del Almacén de Datos, el cubo de datos y el modelo dimensional propuesto.

Finalmente en el Capítulo 3 se detallan las temáticas referidas al modelo de datos físico, la estructura de la base de datos (BD), análisis del rendimiento, pruebas, así como la validación general del Sistema y el análisis de los resultados.

Capítulo I: Fundamentación Teórica

Introducción

En este Capítulo los puntos implicados están enmarcados en la investigación de los portadores energéticos, se abordarán los principales indicadores energéticos por los cuales se miden el consumo de energía y de herramientas existentes para este tipo de control. Se establecerá un análisis de las tecnologías de almacenamiento de datos así como se hará un estudio más profundo sobre los Sistemas de Almacenes de Datos, sus principales metas y los elementos que los componen. Se realizará además un estudio del estado del arte tanto a nivel mundial como nacional sobre las metodologías existentes y las principales herramientas para el desarrollo de los almacenes de datos, así como la justificación de su uso.

1.1 Portadores Energéticos

Son recursos naturales o artificiales que pueden ser procesados o transformados en fuentes de energía según la necesidad del ser humano, esta transformación puede tener varios usos, por ejemplo, como combustible para los vehículos, el calor y la luz de las casas, ejecutar los reproductores de música, cargar las baterías, entre otras propiedades según el fin que se le dé.

1.1.1 Tipos de Portadores Energéticos más Comunes son:

- **Gas Natural:** Es una mezcla de gases compuesta principalmente por metano. Se trata de un gas combustible que proviene de formaciones geológicas, por lo que constituye una fuente de energía no renovable. (Definición.de, 2008).
- **Petróleo:** Líquido natural oleaginoso e inflamable, constituido por una mezcla de hidrocarburos, que se extrae de lechos geológicos continentales o marítimos. Mediante diversas operaciones de destilación y refinado se obtienen de él distintos productos utilizables con fines energéticos o industriales, como la gasolina, la nafta, el queroseno, el gasóleo. (OPEP, 2008).

Para mantener un control de estos portadores energéticos y posteriormente realizar análisis de comportamiento del consumo de estos, es necesaria la estadística, la cual se apoya en variables que la

ayudan en sus cálculos, este tipo de variable se conoce como Indicadores.

1.2 Indicadores

Existen tres tipos de indicadores fundamentales:

- **Los indicadores de proceso:** Se definen como el conjunto de datos obtenidos durante la ejecución del proceso que permiten conocer el comportamiento del mismo y por tanto, predecir su comportamiento futuro en circunstancias similares. (Unión de Asociaciones, 2007).
- **Los indicadores de producto:** Son el conjunto de datos referidos al producto en sí (medidas obtenidas respecto a medidas previstas, por ejemplo) cuyo análisis indica hasta qué punto se ha conseguido el producto que se deseaba. (Unión de Asociaciones, 2007).
- **Los indicadores de servicio:** Igual que los indicadores de producto, son el conjunto de datos referidos al servicio cuyo análisis indica el grado de cumplimiento de los niveles de servicio previamente establecidos. (Unión de Asociaciones, 2007).

1.2.1 Principales Indicadores Energéticos En el Mundo

Indicadores (Organismo Internacional de Energía, 2005):

- **SOC3:** Uso de energía en los hogares por grupo de ingresos y combinación de combustibles utilizados.
- **ECO6:** Intensidades energéticas de la industria.
- **ECO9:** Intensidad energética de los hogares.

1.2.2 Principales Indicadores Energéticos En Cuba

- Comportamiento del consumo de Electricidad por provincias.
- Comportamiento del consumo de Queroseno por provincias.
- Transformación del consumo energético en el hogar.

Tanto en Cuba como en el mundo se han ido desarrollando herramientas que brindan apoyo para el tratamiento y control de la estadística utilizando a los indicadores para un mejor análisis, evitando el

engorroso trabajo con el cúmulo de información que hoy día cada institución genera.

1.3 Herramientas Existentes para el Control de Energía

La utilización de herramientas para este tipo de control ayuda al hombre a ser más exacto a la hora de realizar estas gestiones energéticas, debido a que estos les permiten a no hacer los trabajos manual, también pueden generar gráficas de comportamiento así como que se pueden integrar a sistemas energéticos para la supervisión y control de electricidad, entre otras funciones que a continuación se abordan por herramientas estudiadas.

1.3.1 Algunas de las Herramientas Existentes En el Mundo

- **CECRE:** Se trata del primer centro con un sistema de control de energías renovables del mundo 21/04/2009. Su función principal es supervisar y controlar las plantas de generación renovables, principalmente los parques eólicos, para garantizar de forma segura la integración de su producción en el sistema eléctrico. (cecre, 2009).
- **Power Studio:** Es un *software* potente, sencillo y con un entorno amigable, con el cual se pueden realizar estudios energéticos de alto nivel. Permite una completa supervisión energética de analizadores de redes y un completo control de diferentes magnitudes en el campo del proceso industrial. (Studio, 2008).

1.3.2 Algunas de las Herramientas Existentes En Cuba

- **La Automática para el ahorro energético en edificios: Xenon Light** es un sistema de control automático integral, que atendiendo a los valores de parámetros de temperatura, caudal, seguridad, ocupación, presión y consumo eléctrico, actúa positivamente sobre los espacios ocupados con énfasis en el ahorro energético. La automatización integral de la instalación resultante permite ahorros energéticos que están entre el 25 y el 40 por ciento del consumo energético anterior. (MIC, 2008).
- **Sistema de Monitoreo de Grupos Electrónicos:** Proyecto que posibilita la supervisión de las principales variables de los equipos de generación distribuida, instalados en los emplazamientos de

todo el País, en aras de una mayor eficiencia energética. Actualmente se encuentra instalado en prácticamente todas las baterías de fuel oil y siempre que entra en línea un nuevo emplazamiento es condición fundamental que tenga el sistema instalado. (minbas, 2008).

1.3.3 Análisis de las Soluciones Existentes

A pesar que las soluciones ya mencionadas con anterioridad permiten controlar y manejar operaciones de los portadores energéticos y además cuentan con excelentes funciones, solo fueron creadas para funciones específicas, pues no posibilitan realizar reportes continuos, ni análisis de datos almacenados. Debido a esto nada de lo existente es factible, ya que se requiere de la construcción de una solución que responda a todas las necesidades del departamento de portadores energéticos de la ONE y por supuesto, para ello es necesario hacer un estudio tanto de tecnologías como de herramientas para seleccionar las adecuadas para la solución a desarrollar. A continuación se hace un estudio de este tipo de tecnología.

1.4 Tecnologías de Almacenamiento de Datos

Desde el surgimiento de las tecnologías de almacenamiento de datos, estas se han ido convirtiendo en una herramienta fundamental para el control y manejo de operaciones, los documentos Excel, XML entre muchos, que aunque hoy día aún se utilizan, ya sea como un simple modo de guardar información o unidas a otras tecnologías resolviendo problemas de almacenamiento y tratamiento de datos, como por ejemplo la utilización de los XML en multimedia y sitios web, que se usan como Bases de Datos, entre otros usos que se le dan a estas tecnologías de almacenamiento. A continuación se realizará un estudio de algunas de estas tecnologías.

1.4.1 Base de Datos

Una base de datos es un conjunto de datos que pertenecen al mismo contexto almacenados sistemáticamente para su posterior uso. En este sentido, una biblioteca puede considerarse una base de datos compuesta en su mayoría por documentos y textos impresos en papel e indexados para su consulta. (Artículo 55, 2008).

Existen varios tipos de Bases de datos:

- **Bases de datos estáticas:** Estas son bases de datos de sólo lectura, utilizadas primordialmente para almacenar datos históricos que posteriormente se pueden utilizar para estudiar el comportamiento de un conjunto de datos a través del tiempo, realizar proyecciones y tomar decisiones. (Artículo 55, 2008).
- **Bases de datos dinámicas:** Estas son bases de datos donde la información almacenada se modifica con el tiempo, permitiendo operaciones como actualización y adición de datos, además de las operaciones fundamentales de consulta. Un ejemplo de esto puede ser la base de datos utilizada en un sistema de información de una tienda, una farmacia, un video-club, etc. (Artículo 55, 2008).

1.4.2 Almacenes de Datos (*DataWareHouse*)

Desde su creación en la década de los 90 hasta la fecha, la tecnología de almacenamiento se ido posicionando como la vía más acertada para la realización de análisis de información histórica. Se ha convertido en una potente herramienta para la recuperación efectiva de las más complejas consultas, además de servir como base para la toma de decisiones.

Los productos *DataWareHouse* han nacido para resolver problemas de análisis de grandes masas de información, en empresas donde una pequeña diferencia en el valor de una variable, puede afectar la cuenta de resultado con unas diferencias de millones de dólares. Incluye tecnologías y herramientas muy diversas, no es un producto que se pueda comprar en el mercado, hay que construirlo, es una colección de datos con características tales como. (Inmon y otros, 2005):

- **Orientada a la materia:** Los sistemas operacionales son clasificados dentro de las compañías según las funcionalidades que realicen, por ejemplo, en una compañía de seguro podrían ser: tratamientos de autos, vida, salud, entre otros. Los Almacenes de Datos se diseñan orientados a los aspectos que son de interés para la organización. Temas como ventas, clientes, vendedores constituyen ejemplos de estos aspectos. Cada compañía posee un conjunto único de temas a analizar. Esta información generalmente es almacenada en fuentes operacionales pero orientadas a un programa específico y dan respuesta a este operativamente.
- **Integrado:** De todos los aspectos de la tecnología *warehousing*, la integración es lo más relevante. Los datos, en el almacén de datos, se alimentan de múltiples y dispares fuentes. Estas fuentes

almacenan la información con el formato que es entendido por el sistema que los va a utilizar lo que evidencia claramente las disímiles variantes que se puede almacenar una misma información.

- **No volátil:** Dentro de la teoría de *DataWarehouse* está definido que la información que es almacenada no se puede modificar directamente. Las únicas dos acciones que se permiten residen en la carga inicial de los datos y el acceso a la información. Todos los procesos establecidos dentro del concepto gestión como son inserción, modificación y actualización se realizan en el ambiente operacional, es decir, todo error que se almacene en el Mercado de Datos se corrige en la fuente de donde vino y se realiza nuevamente el proceso de carga de los datos.
- **Varían con el tiempo:** Tiempo variante significa que cada unidad de dato almacenada es exacta a partir de un momento en el tiempo. En algunos casos el registro está sellado en el tiempo y en otros posee una fecha de transacción, pero en cualquier caso existe una forma de mostrar el momento en el tiempo cuando se realizó el registro.

Ventajas de los *DataWarehouse*:

- Transforma datos orientados a las aplicaciones en información orientada a la toma de decisiones.
- Permite un análisis inmediato de los resultados de compras, ventas.
- Agilidad en el control de stocks.
- Capacidad de analizar y explorar las diferentes aéreas de trabajo.
- Relación total con el cliente.
- Facilidades en la gestión y análisis de recursos.
- Conexionar departamentos empresariales (que antes formaban islas).
- Reaccionar rápidamente a cambio del mercado.

Desventajas de los *DataWarehouse*:

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

1.4.3 Mercados de Datos (*DataMart*)

Los Mercados de Datos, también conocidos como *DataMart*, son *DataWarehouse* orientados a temas específicos o aplicaciones específicas que contienen datos de sólo una línea del negocio. La mayor

diferencia entre ellos es el ámbito de la información que contienen debido a que en los *DataMart* son más pequeños y los datos se obtienen de un menor número de fuentes y comúnmente el tiempo de desarrollo es menor. (Hobbs y otros, 2005). Un *DataMart* es una alternativa de solución al igual que *DataWarehouse* a los problemas antes planteados porque el diseño y construcción son similares, además de poseer una secuencia común. La diferencia entre estas dos estructuras se basa principalmente en que los *DataMart* están enfocados en un área de negocio específica, mientras que un *DataWarehouse* entrega información a nivel corporativo. (Peñaloza, 2008). Se puede decir que los *DataMart* están conectados con la arquitectura de los *DataWarehouse* en su forma más simple y que representan los datos de un sólo proceso del negocio a la vez. (Kimball y otros, 2002). Los *DataMart* se caracterizan por disponer una estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos del departamento al cual está aplicado.

1.4.4 Justificación de la Tecnología de Almacenamiento a Utilizar

Se selecciona como tecnología de almacenamiento a los *DataWarehouse* debido a que los sistemas transaccionales de almacenamiento usan estructuras normalizadas, en las cuales se optimizan las inserciones y actualizaciones de datos e incluso algunas selecciones de los mismos, pero es menos probable que este tipo de sistema se organice de forma tal que produzca reportes eficientes para datos resumidos con cierta jerarquía, por tanto para dar solución a esto se podría utilizar un *DataMart* por sus disímiles características, siendo este un subconjunto de datos que contiene un *DataWarehouse*. Pero debido a que los *DataMart* son enfocados a un área de negocio específica y enmarcada por departamentos, en un área donde se engloban un conjunto de estos, para separar las temáticas por su clasificación, con grandes volúmenes de información, no sería factible como solución para el problema planteado. Además la temática a tratar aunque es específica, es una sola y con poca información, sin necesidad de la utilización de un *DataMart*, pero se deja planteado para un futuro si esta temática aumenta o urge la necesidad de dividirla. Debido a esto, debería usarse un *DataWarehouse* ya que capta los datos relevantes de fuentes existentes y las combina en estructuras que ya han sido optimizadas para las selecciones, así como las disímiles facilidades que nos ofrece, tales como que organizan los datos en el almacén en modelos estructurales, para brindar a la gerencia un grupo de informaciones organizadas en cubos multidimensionales que les permite profundizar en el análisis de la información y ver su variación

en el tiempo, entre otras características, propiedades, ventajas y diferentes funciones ya descritas con anterioridad.

Al ser definido el almacén de datos como la tecnología de almacenamiento adecuada para la solución a desarrollar, se debe seleccionar una metodología que defina los pasos necesarios que sirvan de guía de cómo llevar a cabo este tipo de desarrollo de software.

1.5 Metodologías para el desarrollo

Para la creación de cualquier tipo de producto informático abarcador, se debería seguir pasos, es ahí donde la metodología brinda un papel fundamental porque se puede categorizar como el conjunto de métodos que se rigen en una investigación. Además es la ciencia que estudia los métodos del conocimiento. Se refiere a los pasos, métodos o procedimientos que se siguen en una investigación para alcanzar los objetivos planteados en dicha investigación. En múltiples disciplinas existen diferentes enfoques para abordar un mismo concepto o problema. La existencia de dichas perspectivas enriquece la propia disciplina. Aplicando dicho principio a las diferentes áreas del conocimiento, se impulsa considerablemente el avance de la ciencia.

El diseño de un Almacén de Datos, como disciplina que ha alcanzado ya un grado de madurez considerable a lo largo de estos años, también presenta diferentes enfoques. En esta tecnología se ha destacado un conjunto de metodologías que definen y guían todo el ciclo de vida del desarrollo concreto. Existen dos criterios bien identificados y que han marcado claramente su tendencia sirviéndole de guía a la comunidad mundial en cuanto al diseño de Almacenes de Datos.

- **Inmon:** Afirma que la creación de una base de datos relacional con una leve normalización es la que nutre los mercados de datos. Por lo que no se crean los *DataMart* directamente desde el sistema de Procesamiento Analítico en Línea (OLTP) a través de un área de ensayo. En lugar de ello, se crean a partir de la arquitectura relacional de los datos corporativos.
- **Kimball:** Divide el mundo de Inteligencia de Negocio (*Business Intelligence*, BI) entre el hecho y las dimensiones es muy eficaz y conduce a una solución completa en una cantidad muy pequeña de tiempo. Además, la técnica de Kimball tiene una gran cantidad de documentación y se puede encontrar una respuesta a casi todas las preguntas que se posean. Entre sus características

principales es que su arquitectura es ascendente (bottom-up) debido a que plantea que se debe crear por cada departamento un conjunto de mercados de datos independientes orientados a los temas que estén relacionados con él. Y “*El Almacén de Datos es la unión de todos los Mercados de Datos de una entidad*”. (Curto, 2008).

- **Hefesto:** Entre sus principales directrices plantea que la construcción e implementación de un Almacén de Datos puede adaptarse muy bien a cualquier ciclo de vida de desarrollo de software, con la salvedad de que para algunas fases en particular, las acciones que se han de realizar serán muy diferentes. Lo que se debe tener en cuenta, no entrar en la utilización de metodologías que requieran fases extensas de reunión de requerimientos y análisis, fases de desarrollo monolítico que conlleve demasiado tiempo y fases de despliegue muy largas. Lo que se busca, es entregar una primera implementación que satisfaga una parte de las necesidades, para demostrar las ventajas del *DataWareHouse* y motivar a los usuarios. (Bernabeu, 2007).
- **Metodología para el Diseño Conceptual de Almacenes de Datos:** Esta última es presentada en la Tesis de Doctorado de Leopoldo Zenaido Zepeda Sánchez. Aporta como aspecto novedoso con respecto a las anteriores la incorporación de los Casos de Uso para guiar el proceso de desarrollo, al mismo tiempo define una serie de transformaciones para llevar desde un diagrama relacional a uno dimensional y así obtener las estructuras que conformarán el Repositorio de Datos. (Zepeda Sánchez, 2008).

1.5.1 Justificación de la Metodología a Utilizar

En la Oficina Nacional de Estadísticas por su papel como órgano rector en materia estadística en Cuba amerita el uso de una metodología robusta y madura que garantice una mejor integración de la información que disponen. Del conjunto de metodologías existentes para el desarrollo de un Almacén de Datos la decisión fue adecuar, la mundialmente conocida, metodología de Kimball, adaptándola, a la realidad de la Universidad de las Ciencias Informáticas (UCI). Esta adaptación fue lograda gracias a los profesionales del grupo de desarrollo del Centro de Tecnologías de Gestión de Datos (DATEC). Los cuales toman a esta metodología como base ya que la misma brinda un conjunto de facilidades y elementos tales como que crea los conceptos de Hechos y Dimensiones, lo que indudablemente es muy eficaz en el proceso de la toma de decisiones y proporciona mayor agilidad en el proceso de desarrollo.

Propone ir construyendo el Almacén de Datos a través de la construcción de los Mercados de Datos departamentales, lo que constituye una estrategia buena y coincide con la división lógica de las empresas, entidades y organismos. Existe abundante documentación sobre la misma, la respuesta a todas las dudas y preguntas que puedan surgir se pueden encontrar en la web, a través de los servicios que brindan el grupo creador de la metodología. Es una metodología madura y reconocida por el resto de la comunidad dedicada al tema. Tiene bien definidas las etapas, actividades, artefactos y roles. Como complemento a la misma y fortaleciendo la etapa del levantamiento de requisitos; se tomó lo planteado por Leopoldo Zenaído Zepeda Sánchez en su Tesis de Doctorado, orientando así el trabajo a los Casos de Uso y se logra estar más alineado con las tendencias y normas de la Universidad.

Al ser seleccionada la metodología de desarrollo para soluciones de almacenes de datos, es necesario definir el tipo de diseño de un modelo de datos que se adapte a la tecnología de almacenamiento seleccionada, para lograr una mejor y más cómoda organización de los datos.

1.6 Modelos de Datos

Un modelo de datos es “un conjunto de conceptos, reglas y convenciones que nos permiten describir y en ocasiones manipular los datos de un cierto mundo real que deseamos almacenar en la base de datos”. Al producto del modelo de datos se le llama esquema (descripción de la estructura de la base de datos) y a los datos en concreto almacenados en la base de datos en ese momento. (Duque, 2008).

Existen varios modelos de BD entre los cuales se encuentra:

1.6.1 Modelo Entidad-Relación

Un diagrama o modelo entidad-relación es un lenguaje para el modelado de datos de un sistema de información. Estos modelos expresan entidades más relevantes para el sistema, sus inter-relaciones y propiedades. Trabajan dividiendo los datos en muchas entidades discretas donde cada una se convierte en una tabla física en la base de datos operacional. Los sistemas de información que se realizan bajo estas directrices comúnmente se denominan sistemas OLTP. Su principal función es reflejar el estado y funcionamiento de las empresas mediante el registro de las operaciones que realizan diariamente. Los modelos entidad-relación no son recomendables para el diseño de los almacenes de datos debido a que

no garantizan la recuperación óptima del gran cúmulo de información que se almacena. Además estos diagramas tienden a resultar en un diseño normalizado mientras que en un almacén de datos este aspecto no es un requisito a tener muy en cuenta. (Hobbs y otros, 2005).

1.6.2 Modelo Dimensional

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

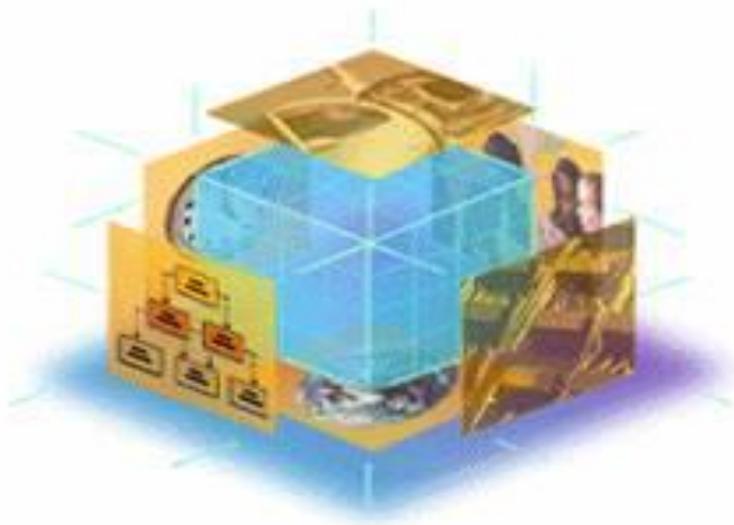


Figura 1: Estructura de un Cubo OLAP (Microsoft Dynamics, 2008).

Para la materialización física de este tipo de modelo se utiliza comúnmente la propuesta realizada por Ralph Kimball llamada “esquema estrella” que consiste en una tabla central denominada “tabla de hechos” y un conjunto de pequeñas tablas, llamadas “dimensiones”, que se relacionan a esta tabla central. Se le

denomina estrella por su similitud con una estrella natural debido a que las dimensiones poseen entre sus relaciones una con la tabla de hechos, ver Figura 2.

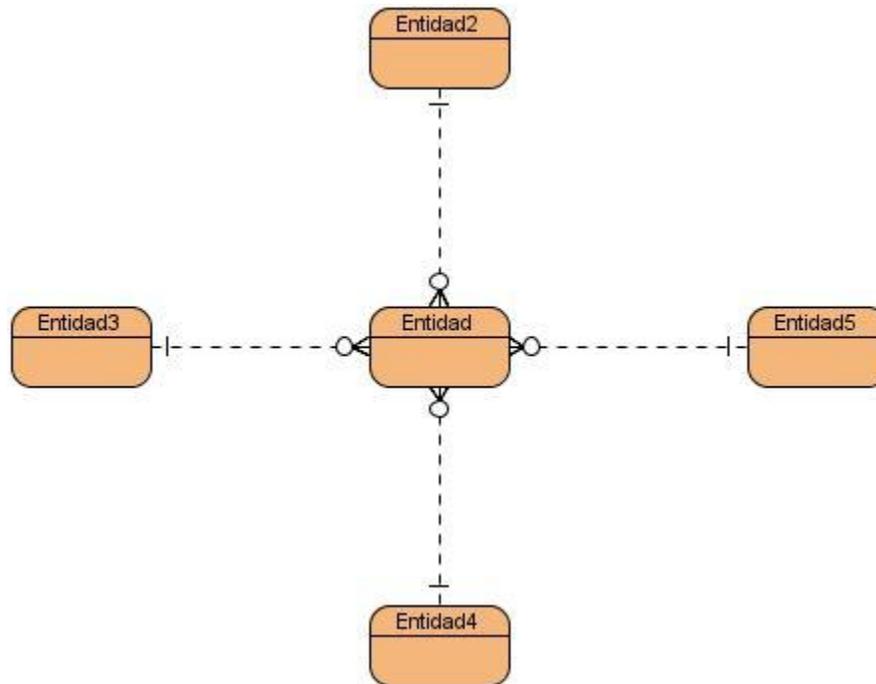


Figura 2: Representación del Esquema Estrella.

Existen otras estructuras que surgen por modificaciones realizadas al esquema estrella. En este sentido se tiene el Copo de Nieve (*snowflaked*) que tiene como objetivo primordial su uso para el ahorro de espacio de almacenamiento. Se dice que una dimensión se encuentra “*snowflaked*” cuando los atributos de baja calidad se llevan a tablas separadas. La utilización de este tipo de estructura posee algunas deficiencias debido a que hace las presentaciones más complejas y afecta el rendimiento de la recuperación de las consultas. Para uso similar al del Copo de Nieve se puede utilizar las Subdimensiones pero sólo es recomendable utilizarlas cuando existe un conjunto de atributos, dentro de las dimensiones, que son necesarios aislar.

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

contiene los valores del negocio. (Ponniah, 2001)

Debido a las características anteriormente mencionadas el modelo dimensional es escogido como modelo de datos a utilizar para el desarrollo de la aplicación.

- **Tablas de Hechos:** La tabla de hechos es la tabla primaria en el modelo dimensional donde el rendimiento de las mediciones numéricas del negocio es almacenado. (Kimball y otros, 2002). Generalmente cada tabla de hecho define un mercado de datos determinado debido a que en ellas se almacena la información concerniente al tema en cuestión, ejemplo ventas, clientes, vendedores.
- **Tablas de Dimensiones:** Las tablas de dimensiones son las compañeras integrales de las tablas de hechos, ellas contienen la descripción textual del negocio. En el modelo dimensional, las tablas de dimensiones poseen varios atributos que en su conjunto definen una fila en la tabla de dimensión. (Kimball y otros, 2002).

Partiendo del estudio realizado sobre los modelos de bases de datos, se decide que para esta investigación la utilización del Modelo Dimensional es el más eficaz, debido a que se ajusta a las características y necesidades de la solución.

Con el fin de brindar soporte al diseño, construcción, implementación y mantenimiento de los Almacenes de Datos, se han desarrollado diversas herramientas. Que a continuación se hacen referencia.

1.7 Herramientas de Modelado

Se puede definir a las Herramientas CASE como un conjunto de programas y ayudas que dan asistencia a los analistas, ingenieros de software y desarrolladores, durante todos los pasos del Ciclo de Vida de desarrollo de un Software. CASE es también definido como el Conjunto de métodos, utilidades y técnicas que facilitan el mejoramiento del ciclo de vida del desarrollo de sistemas de información, completamente o en alguna de sus fases. A continuación se abordarán algunas de estas herramientas:

- **Rational Rose:** Es una herramienta software para el modelado de sistemas software. Permite Especificar, Analizar, Diseñar el sistema antes de Codificarlo. Mantiene la consistencia de los modelos del sistema software, chequea la sintaxis UML, permite generar documentación automáticamente, código a partir de los Modelos y nos posibilita realizar la ingeniería Inversa (crear modelo a partir código). (Grupo soluciones, 2009).

- **Visual Paradigm:** Es una herramienta UML fácil de usar que soporta la última notación UML 2.1, nos permite realizar la ingeniería inversa, generar código, importar desde Rational Rose, exportación/importación XMI, nos posibilita generar informes, editar figuras. Entre sus nuevas características incluyen el modelado colaborativo con Subversion, interoperabilidad con modelos UML2 a través de XMI. (Company Headquarters, 2009).

Características:

- Diseño centrado en casos de uso y enfocado al negocio que genera un software de mayor calidad.
- Uso de un lenguaje estándar común a todo el equipo de desarrollo que facilita la comunicación.
- Disponibilidad en múltiples plataformas.
- Contiene facilidades para redactar Especificaciones de Casos de Uso del Sistema.
- Sincronización entre Diagramas de Entidad Relación y Diagramas de Clases.
- Generación de documentos.
- Integración con distintos Ambientes de Desarrollo Integrados (IDE).

1.7.1 Justificación de la Herramienta CASE a Utilizar

Se selecciona al Visual Paradigm como herramienta case por su integración a UML además de portar con las características de ser multiplataforma, amigable en su uso y poseer interoperabilidad con otras aplicaciones e integración con distintos IDE. También por sus disímiles características ya mencionadas con anterioridad, no sin mencionar que como herramienta de su tipo se apoya en un lenguaje. Lenguaje Unificado de Modelado (UML), es un lenguaje gráfico para visualizar, especificar, construir y documentar un sistema. UML ofrece un estándar para describir un "plano" del sistema (modelo), incluyendo aspectos conceptuales tales como procesos de negocio, funciones del sistema y aspectos concretos como expresiones de lenguajes de programación, esquemas de bases de datos y componentes reutilizables. Es importante resaltar que UML nos permite especificar o describir métodos o procesos. Se utiliza para definir un sistema, detallar los artefactos en el sistema, documentar y construir. En otras palabras, es el lenguaje en el que está descrito el modelo.

Ya teniendo la herramienta con la que se modelará la investigación, se hace necesaria la selección de un

gestor de base de datos que cumpla con las condiciones especificadas para darle respuesta a la solución.

1.8 Gestores de Base de Datos

Han surgido compañías que han marcado hitos en este sentido, un ejemplo de ellas es el gigante Oracle, la cual en la actualidad es la número uno en el desarrollo de aplicaciones de este tipo, tanto para el mantenimiento de las bases de datos en función de las mejoras de las consultas, la optimización y sistemas estratégicos de negocios empresariales enfocados a la toma de decisiones a nivel empresarial.

Con su producto:

- **Oracle:** Ofrece un rápido, fiable y seguro intercambio de información, análisis y extracción de datos a bajo costo, y redes escalables. Obtiene un rendimiento extremo y la escalabilidad de los almacenes de datos, gestiona la carga e integración de datos, mejora el rendimiento del almacén de datos, disponibilidad y capacidad de administrar grandes tablas de partición, encuentra patrones ocultos y puntos de vistas en su almacén de datos. Lidera el mercado de almacenes de datos. (Oracle, 2010).
- **PostgreSQL:** Es un sistema de gestión de bases de datos objeto-relacional (*ORDBMS*) basado en el proyecto POSTGRES de la Universidad de Berkeley. Es una derivación libre (*OpenSource*) de este proyecto y utiliza el lenguaje SQL92/SQL99. Fue el pionero en muchos de los conceptos existentes en el sistema objeto-relacional actual, incluido más tarde en otros sistemas de gestión comerciales. Incluye características de la orientación a objetos, como puede ser: la herencia, tipos de datos, funciones, restricciones, disparadores, reglas e integridad transaccional. A pesar de esto, no es un sistema de gestión de bases de datos puramente orientado a objetos. (Group, 1996-2010.).

1.8.1 Justificación del Gestor de Base de Datos a Utilizar

Actualmente la ONE está exportando todos sus dispositivos de almacenamiento hacia plataforma PostgreSQL por lo que queda seleccionado como **Sistema Gestor de Base de Datos el PostgreSQL**. En este sentido la versión que se utilizará es la 8.4 por ser lo suficientemente estable, segura y la más actualizada. Esta decisión ha sido previamente colegiada por parte del cliente final debido a que dentro

de sus políticas de migración se encuentran las de llevar a todas sus bases de datos hacia dicha plataforma.

- **PostgreSQL8.4:** Esta nueva versión contiene una gran cantidad de mejoras para la administración, consulta y programación de bases de datos PostgreSQL sea más fácil que nunca. Se les fue añadido más de doscientas mejoras en todos los aspectos de la funcionalidad de base de datos, ayudando a todos los usuarios de PostgreSQL. Muchos de los cambios en PostgreSQL 8.4 son nuevas o mejoradas herramientas de administración y control de los comandos. Entre las mejoras más populares se encuentra que, hace más fácil el análisis de datos a través de la avanzada ANSI SQL2003 características de ventanas de funciones, expresiones de tabla común y consultas recursivas. Mejoras a los procedimientos almacenados, tales como los parámetros por defecto y los parámetros variables en el tiempo, hacen la programación de servidor de base de datos más simple y más compacta. Por supuesto, también permite restaurar la base de datos paralelamente, acelerar la recuperación de copia de seguridad hasta 8 veces. Los permisos de columna, lo que permite un control más granular de los datos sensibles. Las actualizaciones a través de `pg_migrator` (beta), que permite actualizaciones de 8,3 a 8,4, sin períodos de inactividad largos. Nuevas herramientas de seguimiento de consultas, a los administradores le dan más visibilidad a la hora de realizar estas operaciones. Gastos generales reducidos en gran medida vacíos a través del mapa de visibilidad. Nuevos instrumentos de control para las consultas en curso, la carga de consulta y bloqueos.

Al ser seleccionado como gestor de Base de Datos el PostgreSQL se hace necesario buscar una herramienta para la administración de este, por sus características es escogido al `pgAdminIII` 1.10, ya que es una herramienta de código abierto para la administración de bases de datos PostgreSQL. Incluye:

- Interfaz administrativa gráfica.
- Herramienta de consulta SQL (con un EXPLAIN gráfico).
- Editor de código *procedural*.
- Agente de planificación *SQL/Shell/batch*.
- Administración de *Slony-I*.

`PgAdminIII` se diseña para responder a las necesidades de la mayoría de los usuarios, desde escribir simples consultas SQL hasta desarrollar bases de datos complejas.

La interfaz gráfica soporta todas las características de PostgreSQL y hace simple la administración. Está

disponible en más de una docena de lenguajes y para varios sistemas operativos, incluyendo Microsoft Windows, Linux, FreeBSD, Mac OSX y Solaris. PgAdminIII soporta versiones de servidores 7.3 y superiores. Versiones anteriores a 7.3 deben usar pgAdminII. (ArPUG, 2009).

Al ser definido el gestor de Base de Datos que se utilizarán en el proceso de desarrollo de la tecnología de almacenamiento de datos seleccionada, se da paso a la selección del modo de almacenamiento de datos que se aborda a continuación.

1.9 Modos de Almacenamiento de Datos

Existen tres modelos para el proceso analítico en línea (OLAP) de la información (ROLAP, MOLAP y HOLAP). El proceso de análisis se realiza de igual forma lo que varía en uno y otro caso en la metodología de almacenamiento. La forma de almacenamiento es crítica para garantizar la velocidad de recuperación de la información, las zonas de ubicación de las agregaciones y el procesamiento de los datos en general.

1.9.1 Procesamiento Analítico Relacional (ROLAP)

El modelo ROLAP es usado fundamentalmente sobre información que no se consulta frecuentemente debido a que no es muy óptimo en este sentido, por ejemplo información histórica de muchos años de antigüedad. Los datos son almacenados en filas y columnas de forma relacional. Este modelo presenta los datos a los usuarios en forma de dimensiones de negocio. Con el fin de ocultar las estructuras de almacenamiento y presentar los datos dimensionalmente es creada la semántica de las etiquetas de los metadatos. Estos metadatos también son almacenados en tablas relacionales. (Sierra, 2009).

1.9.2 Procesamiento Analítico Multidimensional (MOLAP)

Aquí las estructuras de los datos están fijas para que la lógica al procesar la información pueda estar basada en métodos bien definidos para establecer las coordenadas del almacenamiento de los datos. Almacena los datos dimensionalmente a diferencia del ROLAP. Para realizar el acceso a la información almacenada de forma más rápida y efectiva utilizándose en depósito donde el tiempo en la velocidad de respuesta es crítico, las estructuras de almacenamiento se organizan en grandes arreglos dimensionales

que son una copia de la fuente de datos y persisten físicamente en la misma estación de trabajo donde está instalada la herramienta *DataWareHouse*. (Sierra, 2009).

1.9.3 Procesamiento Analítico Híbrido (HOLAP)

Como su nombre lo indica, es un híbrido entre los métodos ROLAP y MOLAP. Permite almacenar una parte de los datos como en un sistema MOLAP y el resto como en uno ROLAP. (Sierra, 2009).

Además posee dos tipos de **aprisionamiento**:

- **Vertical:** Almacena las agregaciones como un MOLAP para mejorar la velocidad de las consultas y los datos se detallan en ROLAP para optimizar el tiempo en que se procesa el cubo.
- **Horizontal:** En este modo se pueden almacenar algunos cubos en MOLAP y otros en ROLAP, así como una sección de los datos, normalmente los más recientes (por ejemplo peticionando por la dimensión tiempo) en modo MOLAP para mejorar la velocidad de las consultas y los datos más antiguos en ROLAP.

1.9.5 Justificación del Modo de Almacenamiento Seleccionado

Se escoge como modo de almacenamiento de datos al Procesamiento Analítico Relacional (ROLAP) debido a que posee un ambiente conocido y disponibilidad de herramientas para su desarrollo lo cual facilita su puesta en práctica para desarrollar la solución. Además de que el tipo de gestor de BD a utilizar exige que sea este el modo de almacenamiento a utilizar.

Teniendo ya definido el gestor de base de datos que se utilizará y el modo de almacenamiento para desarrollar la investigación, hay que tener en cuenta cómo se integrarán los datos al almacén, para esta integración se emplean herramientas las cuales son utilizadas en los procesos de Extracción Transformación y Carga (ETL).

1.10 Herramientas para el proceso ETL

En la actualidad han aparecido nuevas posibilidades de adquisición de soluciones para el proceso de ETL. Pues actualmente son diversas y numerosas las herramientas de las cuales resultaría provechoso un estudio. Entre las herramientas más conocidas encontramos; *Power Center*, una plataforma de integración

de datos lanzada por *Informatic Corporation*, que incluye la construcción de almacenes de datos de terabytes. Otras de las más conocidas son mencionadas a continuación:

- **Talend Open Studio:** Talend es un proveedor de productos para la integración de datos, compañía de propiedad privada y respaldada por capital arriesgado. Esta herramienta es un producto generador de código, que permite ser analizado cada vez que se requiera, esto significa que evita la concepción de ser una caja negra seguida por otros desarrolladores. Soporta las siguientes plataformas: Solaris, MAC, Windows, Red Hat Enterprise Linux, Linux (permite arquitecturas de 32 y de 64 bits). Entre sus principales ventajas, se encuentra el foro de la compañía que proporciona soporte de manera gratuita, además la herramienta permite equilibrar la carga entre el servidor de procesamiento de Talend, grupo o red y el origen o destino de las bases de datos en escena. Otra ventaja es que cuenta con una interfaz ETL para la importación de los metadatos, la configuración, vinculación de los componentes y generación de código, lo cual proporciona ganancias de productividad para los desarrolladores. Las desventajas que ofrece esta herramienta no son muchas, la más significativa es el no aseguramiento de la calidad de los datos, aunque esta función puede ser cumplida por una tercera parte. (Talend, 2009).
- **Spoon de Pentaho Data Integration:** Spoon de Pentaho Data Integration es una de las herramientas ETL código abierto más populares, la cual permite diseñar de forma gráfica la transformación ETL. Reúne un conjunto de componentes que permiten modelar y ejecutar transformaciones sobre flujos de datos. Incorpora operaciones de transformación, así como funciones para operar con los campos en el flujo de datos, renombrando, calculando campos en función de otros, correlacionando valores y realizando búsquedas auxiliares en bases de datos. Por otra parte ofrece, soporte para operaciones de dimensiones lentamente cambiantes, permite ejecutar código javascript dentro de las transformaciones e incorpora un evaluador de expresiones regulares. Entre las tareas que se pueden incorporar están la copia y eliminación de ficheros, así como su transportación usando el Protocolo de Transferencia de Archivos (FTP), además se pueden descompactar ficheros. Esta herramienta es fácil de usar: es model-driven, por lo que no es necesario prácticamente incorporar código, aunque el usuario tiene la posibilidad de programar sus propias transformaciones y desarrollar un proceso que consiste en declarar el flujo en una interfaz gráfica. Brindando la posibilidad de copiar y leer del mismo fichero en paralelo, permitiendo maximizar la capacidad de entrada/salida en el entorno ETL. (Pentaho, 2009).

1.10.1 Justificación de la Herramienta de Integración a Utilizar

Para apoyar el proceso ETL en la integración de los datos al Almacén, de las herramientas ya mencionadas se toma como la factible a **Spoon de Pentaho Data Integration**:

- **Spoon de Pentaho Data Integration v 3.1.0 (PDI)**: Es de formato abierto y de fácil lectura para los Lenguajes de Marcas Extensibles (XML) que recogen transformaciones, tareas programadas y un repositorio relacional de metadatos ETL. Es aplicable a diversos tipos de bases de datos (SQL server, PostgreSQL, MySQL, Microsoft Access). Posee facilidad para la importación y exportación de datos de un formato a otro cualquiera. Su principal fortaleza es la posibilidad que brinda de ser extensible mediante pluggins. Alta compatibilidad con la herramienta ETL con gestor PostgreSQL en los llamados Paquetes, donde se ofrecen servicios de importación, exportación, transporte, y transformación de datos. Rendimiento eficiente: Alta velocidad de respuesta de las consultas. El servidor PostgreSQL posee un potente motor de recuperación de los datos y permite la optimización de las complejas consultas enviadas o preparadas desde el PDI. Capacidad del PDI para extraer y cargar datos utilizando el PostgreSQL. Con estas herramientas se tiene la posibilidad de no sólo almacenar los datos al más atómico de los detalles, sino que también se puedan guardar solamente los agregados necesarios.

Para lograr una buena visualización de los datos ya integrados al almacén por la herramienta seleccionada, es necesario realizar un estudio de los software utilizados para esto, esta visualización forma parte de los procesos de Inteligencia del Negocio (BI). A continuación se realizará un estudio de herramientas utilizadas en estos procesos.

1.11 Herramientas para el proceso BI

En la actualidad han aparecido nuevas posibilidades de adquisición de soluciones de BI. Existen productos de software *OpenSource* disponibles para cualquier función de una plataforma de toma de decisiones o BI y que proporcionan un nivel de uso, de capacidad, de carga y robustez compatibles con necesidades críticas. En el plano *OpenSource* la herramienta más significativa es Mondrian, la cual es una de las aplicaciones más importantes de la Suite Pentaho BI.

- **Pentaho Mondrian**: Es un servidor OLAP *opensource* que gestiona la comunicación entre una

aplicación OLAP y la base de datos con los datos fuente. Es desarrollado en Java/Servlets/JSPs que permite ser instalado en servidores de aplicaciones como JBoss. Permite crear cubos de información, dichos cubos se componen de archivos XML y en ellos se definen las Dimensiones y las conexiones de los datos. Los archivos XML por lo general son complejos de realizar manualmente por lo que es común utilizar herramientas gráficas para realizar la edición de estos. El núcleo del servidor Mondrian es similar a la Conexión del Java a la BD (JDBC) pero exclusivo para OLAP. Proporciona la conexión a la base de datos y ejecuta las sentencias SQL. (Mining, 2008)

Los esquemas Mondrian se alojan en un archivo XML. Este esquema puede ser editado de forma manual, o utilizando las herramientas Workbench o Cube Designer de Mondrian.

- **Pentaho Workbench o Cube Designer de Mondrian:** Es un entorno visual para el desarrollo y prueba de cubos OLAP Mondrian. Si bien la definición del XML para esquemas Mondrian no es extremadamente compleja, en la práctica resulta engorroso recordar cada uno de los elementos junto a sus atributos y sub-elementos. Con esta aplicación, se puede configurar una conexión JDBC como el modelo físico, para luego elaborar el esquema lógico de manera simple y efectiva. Para ello el entorno ofrece un editor de esquemas con la fuente de datos subyacente para su validación. Permite la ejecución de consultas de Expresiones Multidimensionales (MDX) contra el esquema y la base de datos y la navegación por la base de datos subyacente. (Mining, 2008).

1.11.1 Justificación de las Herramientas para los Procesos de BI

Como solución en la plataforma de la Suite de Pentaho BI, **como motor OLAP se tomó como herramienta a Pentaho OLAP Mondrian:**

- **Pentaho OLAP Mondrian 3.1.1:** Entre sus principales características se encuentra la facilidad para el análisis de grandes volúmenes de información que se encuentren almacenados en bases de datos. Soporta el lenguaje *Microsoft's Multidimensional Expressions (MDX)*. También soporta los APIs: Java OLAP (JOLAP) y *XML for Analyss application programming*. Permiten realizar consultas al data *warehouses* y permite que los resultados sean presentados mediante un navegador de modo que el usuario pueda realizar las actividades típicas de navegación.

Para evitar el trabajo de forma manual con los esquemas que se generan en Mondrian, **se selecciona como herramienta de edición a Workbench en su versión 3.1.1.**

- **Pentaho Workbench 3.1.1:** Esta es una herramienta para el desarrollo del esquema del modelo estrella en XML desarrollada en Java. Este programa recientemente publicado (2007) entrega todas las facilidades para poder realizar el modelo lógico del cubo OLAP al cual se le realizarán las consultas. Este programa se conecta directamente con la base de datos para así poder diseñar los cubos OLAP que se requieren para que el usuario final pueda visualizar los indicadores en el *DataMart*. Luego, el archivo generado se utiliza para definir la estructura de cubo en Jasper. (Mining, 2008).

Otros aspectos a tener en cuenta son:

Alta compatibilidad con las herramientas ETL, en este caso Pentaho Data Integration, del gestor PostgreSQL en los llamados Paquetes, donde se ofrecen servicios de importación, exportación, transporte, y transformación de datos.

Rendimiento eficiente: Altísima velocidad de respuesta de las consultas. El servidor PostgreSQL posee un potente motor de recuperación de los datos y permite la optimización de las complejas consultas y reportes dinámicos.

Posibilidad de integración de múltiples orígenes de datos: Brinda la capacidad de combinar los datos de distintas fuentes, lo cual suele ser una tarea bastante complicada para las personas encargadas de tomar decisiones con esa información sin tener que homogeneizarla en un mismo repositorio.

Soporte de agregados: Con esta herramienta se tiene la posibilidad de no solo almacenar los datos al más atómico de los detalles, sino que se pueden guardar los agregados necesarios.

1.12 Herramientas seleccionadas para la realización de pruebas

Para la realización de las pruebas se seleccionaron las herramientas **Jmeter** y **PgData Generator** debido a que:

JMeter: Es una herramienta de carga para llevar a cabo simulaciones sobre cualquier recurso de software. Inicialmente diseñada para pruebas de estrés en aplicaciones web, hoy en día su arquitectura ha evolucionado no sólo para llevar a cabo pruebas en componentes habilitados en Internet (HTTP), sino

además en Bases de Datos, programas en Perl, requisiciones FTP y prácticamente cualquier otro medio.

Además, posee la capacidad de realizar desde una solicitud sencilla hasta secuencias de requisiciones que permiten diagnosticar el comportamiento de una aplicación en condiciones de producción. En este sentido, simula todas las funcionalidades de un Navegador (“Browser”), o de cualquier otro cliente, siendo capaz de manipular resultados en determinada requisición y reutilizarlos para ser empleados en una nueva secuencia. (Softqanetwork, 2005)

PgData Generator: Es una herramienta sencilla, potente y totalmente personalizable que genera datos para propósitos de prueba base de datos. Gran ventaja de este software es que es capaz de crear una amplia variedad de tablas de prueba y apoya las plantillas definidas por el usuario. Estos datos pueden ser fácilmente insertados en una base de datos. Hay varias maneras de llenar las bases de datos al azar, por una máscara, o con datos de otra tabla. Es importante destacar que la configuración de cada cuadro son controlados separately. DTM Data Generator es un software que puede generar datos para las pruebas base de datos. Un paquete de datos puede contener múltiples tablas, cada una con propias reglas, rango de valores y parámetros, lo que hace uso de la herramienta muy conveniente. Además, DTM Data Generator es capaz de crear las instrucciones SQL para cualquier generador de datos operation. DTM es muy sencillo y funciona muy rápido. Es una herramienta indispensable para cualquier desarrollador o administrador de base de datos que quiere ahorrar tiempo. (Altacracks, 2007).

1.13 Conclusiones del Capítulo

La Oficina Nacional de Estadísticas es una de las entidades que se encuentran en franca migración hacia la independencia tecnológica, debido a esto las herramientas seleccionadas para el desarrollo de la solución son de aquellas que por sus propiedades sean las más factibles y que cumplan con esta migración.

A partir del estudio del arte realizado se concluyó que la utilización de la tecnología de Almacén de Datos, posibilita integrar fuentes de datos muy diversas y ponerlas a disposición de los usuarios, para su consulta de una manera eficaz apoyando el proceso de toma de decisiones. Luego del estudio sobre las diferentes herramientas que brindarán apoyo a las construcción de la solución, se selecciona como sistema gestor de Base de Datos el PostgreSQL en su versión 8.4, para los procesos de BI y ETL se seleccionó el Pentaho Data Integration 3.1.0, PentahoOLAP Mondrian 3.1.1, Pentaho Workbench 3.1.1, como herramienta de modelado al Visual Paradigm apoyándose en el lenguaje UML, como metodología se tomó a el modelo de desarrollo para Almacenes de Datos e Inteligencias de Negocio, así como modelo de datos el Multidimensional y como modo de almacenamiento el ROLAP.

Antes de concluir es necesario hablar de la Herramienta de control de versiones para mantener salvadas de los documentos, versiones de productos y poder acceder a ellas mediante la red, dado que esta herramienta (Subversión vs 1.4.5) posibilita que varias personas puedan trabajar con el mismo conjunto de datos lo que fomenta la colaboración entre estas y garantiza la calidad a la par que disminuye el tiempo de trabajo. Además del Tortoise 1.4.5, herramienta cliente del Subversión, la ventaja de este producto es que se instala como una extensión del escritorio de modo que pulsando sobre el botón derecho del mouse, este es capaz de traer documentos y actualizar el repositorio.

Capítulo II: Análisis y Diseño

Introducción

Un Almacén de datos está compuesto por una serie de componentes con responsabilidades específicas que trabajan de manera semi-independiente. Cada uno tributa desde su espacio al desarrollo de una solución que se adapte a las necesidades de los usuarios finales.

Estos componentes están enmarcados dentro de 4 grandes grupos según la metodología a utilizar para el desarrollo de la solución:

- Análisis de la Solución propuesta.
- Diseño dimensional de las estructuras.
- Extracción, Transformación y Carga (ETL) de los datos de la(s) fuente(s).
- Inteligencia de Negocio (BI).

2.1 Análisis de la Solución

En este epígrafe se realiza el análisis necesario para concretar la solución haciendo énfasis en una serie de actividades que lo posibilitan tales como la definición del negocio y sus reglas, el esclarecimiento de el tema de análisis, los roles y permisos además de la especificación de los requisitos funcionales, no funcionales, de información y multidimensionales según las necesidades del usuario lo cual posibilita la elaboración de los casos de uso del sistema.

2.1.1 Definición del Negocio

La ONE constituye el ente rector y coordinador de los temas estadísticos en Cuba. Entre sus objetivos se encuentra el de funcionar como un repositorio central de información, donde se supervisa y controla la gestión estadística del País. Dicha organización también se encarga del control estadístico de sus oficinas en las diferentes regiones y en particular del control estadístico de portadores energéticos. El consumo de estos portadores es divulgado a través de reportes que representan las necesidades de información de los

clientes. La información publicada está estructurada en planillas que son enviadas en una fecha determinada desde cada oficina, ya sea municipal, territorial y la propia sede central.

2.1.2 Tema de Análisis Identificado

La definición del tema de análisis (TA) es uno de los pasos más importantes en el proceso de análisis de un *DataWarehouse*, pues enfoca el desarrollo de la solución en las áreas temáticas presentes en el negocio, facilitando así el cumplimiento de las metas trazadas. Los temas se definieron en función de los diferentes cortes de información que realiza la ONE en correspondencia con las necesidades de la institución.

- **Comportamiento de los Portadores Energéticos** en este TA se identificaron Indicadores que cumplen con la necesidad para la realización de un buen análisis de esta área:
 - Consumo de Portadores Energéticos.

2.1.3 Roles y Permisos

Uno de los objetivos de la ONE, en función de llevar un control sobre el consumo de los portadores energéticos, es contar con una aplicación que permita a las personas encargadas trabajar con la información proveniente de las oficinas regionales. Por lo planteado anteriormente se cuenta con los especialistas que son los encargados del análisis de la información enviada por las oficinas de la ONE, así como que pueden consultar la información y hace uso de ella para los cálculos estadísticos. El administrador es el encargado de los procesos ETL.

Los permisos de acceso y el trabajo con el sistema están definidos como se informa a continuación y los usuarios fuera de esta clasificación no podrán tener permisos de acceso al sistema:

- Analistas: Permiso de consulta sobre los datos del Almacén.
- Administrador: Permiso de consulta, escritura y actualización sobre los datos del Almacén.

2.1.4 Reglas del Negocio

El trabajo con la información enviada por las oficinas de la ONE para el control estadístico, especialmente en el TA a tratar, requiere que:

- La información sea enviada en una fecha establecida por la sede central, para lograr un análisis continuo y sistemático.
- El formato de los reportes sea en Excel, documento el cual la ONE esparcirá por sus oficinas debido a la no existencia de un documento estándar, esta planilla será definida según las necesidades de este TA y contará con los campos necesarios para evitar problemas a la hora de interactuar con el *DataWarehouse*. Este documento además contará con los campos que reflejará el comportamiento según el análisis que se haga, los cuales se explicaran a continuación. Estos Excel forman parte del código fuente del Expediente de Proyecto:
 - **comportamiento_cumplimiento_consumo** (combustible o electricidad): Este tipo comportamiento es calculable por el almacén expresando el cumplimiento o incumplimiento del consumo asignado, necesitando para su composición la resta entre este consumo asignado y el real.
 - **consumo_portadores_energeticos**: Para poder realizar análisis y observar comportamientos más generales de estos portadores energéticos.

2.1.5 Necesidades de los Usuarios

La propuesta de solución está encaminada a resolver un grupo de necesidades existentes en la ONE por parte de sus trabajadores, que básicamente son los potenciales usuarios de la futura aplicación. Debido a lo anteriormente expresado, es necesario y de vital importancia identificar sus necesidades, alineando las mismas a él TA definido. Es por ello que se define el enfoque hacia los siguientes dominios de información:

1. Indicadores sobre el Consumo de Electricidad.
2. Indicadores sobre el Consumo de Combustible.

2.1.6 Requisitos de Información

Los requisitos de información representan toda la información que debe estar disponible en el almacén para su consulta. Constituyen la entrada fundamental para todo el proceso de BI y para futuros reportes bases. A continuación se mostrarán los requisitos de información que fueron identificados durante el proceso de análisis, agrupándolos por dominios informativos.

Combustible:

- Obtener el consumo de Combustible por: oficina, tiempo, vehículo, tipo de combustible, las actividades realizadas.
- Obtener la variación del consumo de Combustible real con respecto al asignado.
- Obtener variación del consumo de Combustible real con respecto a reportes anteriores.

Electricidad:

- Obtener el consumo de Electricidad por: tiempo, oficina, tipo de electricidad.
- Obtener la variación del consumo de Electricidad real con respecto al asignado.
- Obtener variación del real consumo de Electricidad con respecto a reportes anteriores.

Portadores Energéticos:

- Obtener el consumo de Portadores Energéticos por: tiempo, oficina.
- Obtener variación del consumo de Portadores Energéticos real con respecto a reportes anteriores.

2.1.7 Requisitos Multidimensionales

Los requisitos multidimensionales constituyen la entrada fundamental para el diseño de las estructuras del almacén. En su totalidad constituyen los parámetros de entrada y salida de las solicitudes de información de los clientes. La profundización de estos parámetros que constituyen este tipo de requisitos serán abordado en el epígrafe 2.3.10.

2.1.8 Requisitos Funcionales

Los requisitos funcionales son capacidades o condiciones que el sistema debe cumplir para dar respuesta a todos los requisitos de información listados anteriormente. Fueron identificados los siguientes:

- Elaboración de una copia de seguridad de la información almacenada, al repositorio una vez al mes.
- Permitir la carga, extracción y transformación de los datos de las distintas fuentes de información.
- Realizar los modelos donde es almacenada la información enviada por las oficinas.
- Permitir los tipos de gráficos de la información que se obtienen en el proceso de BI.

2.1.9 Requisitos no Funcionales

Son propiedades o cualidades que el producto debe cumplir. Especifican criterios que pueden usarse para juzgar la operación de un sistema en lugar de sus comportamientos específicos. Por tanto, se refieren a todos los requisitos que ni describen información a guardar, ni funciones a realizar, siendo específicamente las propiedades y características que hacen al producto atractivo, usable, rápido y confiable. Estos requisitos pertenecen al código fuente del expediente de proyecto, específicamente se pueden encontrar en el artefacto Especificación de requisitos.

2.1.10 Casos de uso del sistema

Los casos de uso del sistema (CUS) se utilizan para capturar los requisitos de un sistema, proporcionando el espacio de interacción del sistema con el usuario o con otro sistema para lograr un objetivo específico. En el desarrollo de la investigación se determinaron casos de uso informativos y funcionales, que se describen y son especificados en el artefacto Modelo de CU del Sistema:

2.2 Diseño de la Solución

A continuación se abordará todo lo referente al diseño de la propuesta de solución, el cual se convierte en un paso importante para la implementación del sistema, identificándose los puntos necesarios para lograr un buen diseño y que se definen a continuación.

2.2.1 Matriz BUS o Matriz Dimensional

El propósito de la Matriz Dimensional es obtener un modelo lógico inicial, (ver Anexo 1), donde quedan identificadas las tablas hechos, las cuales forman las áreas de análisis del Almacén de Datos y sus dimensiones relacionadas.

Áreas de Análisis (AA) del tema de análisis identificado:

- Consumo de Combustible (**hech_consumo_combustible**).
- Consumo de Electricidad (**hech_consumo_electricidad**).

- Consumo de Portadores Energéticos (**hech_portadores_energéticos**).

Tablas dimensiones:

- Electricidad (**dim_electricidad**).
- Combustible (**dim_combustible**).
- Oficina (**dim_oficina**).
- Temporal (**dim_temporal**).
- Vehículo (**dim_vehiculo**).
- Actividad (**dim_actividad**).

En la Tabla 1 se relacionan las AA identificadas con las dimensiones propuestas con el fin de presentar el Almacén de Datos candidatos del proceso en cuestión.

AA/DIM	Combustible	Electricidad	Oficina	Vehículo	Actividad	Temporal
Consumo Combustible	x	-	x	x	x	x
Consumo Electricidad	-	x	x	-	-	x
Portadores Energéticos	-	-	x	-	-	x

Tabla 1: Matriz BUS Relación Áreas de Análisis y Dimensiones.

2.2.2 Modelo de Datos

Al ser seleccionados los posibles hechos y las dimensiones candidatas para la solución, es importante conocer la composición de ambos y su función como parte del almacén.

1. Tablas dimensiones:

Describen las dimensiones y jerarquías que están relacionadas con el repositorio principal donde se va a almacenar la información, del Modelo de Indicadores Generales, atómicamente. Estas dimensiones se pueden encontrar en el artefacto Especificación de Dimensiones el cual forma parte del Expediente de Proyecto.

2. Tablas de hechos:

Las tablas de hechos son las fuentes de almacenamiento de las medidas numéricas. En este caso se definieron como medidas numéricas los valores que se captan en la propuesta del Excel que son lo concernientes al consumo de los portadores energéticos: Real, Asignado del mes o año que se analiza. La tabla de hechos identificada se describe a continuación:

Las Tablas (hech_consumo_combustible), (hech_consumo_electricidad) y (hech_portadores_energeticos): Son las tablas que servirá como fuente de información principal para la realización de las estructuras que soporten los diversos reportes más comunes de la institución. Siendo estas tablas las que van a residir como repositorio central, para toda la información existente en los Excel estadísticos. Estas tablas son descritas en el artefacto, Especificación de Tablas Hechos.

3. Medidas

Son las variables de salida del Almacén, las cuales pueden estar o no en las tablas hechos. Clasificándose como variable físicas, que no son más que aquellas que si están en las tablas hechos y las calculables no son más que las que el almacén devuelve tras algún análisis realizado. Estas medidas son encontradas en los artefactos, Especificación de Tablas Hechos y en el Excel, Herramienta de Recolección de Información.

2.2.3 Arquitectura de los Componentes del Sistema

Una arquitectura dentro del desarrollo de un software es el diseño estructural de dicho sistema o producto basado en objetivos y restricciones. Específicamente, en la tecnología *warehousing*, es una forma de representar toda la estructura de datos, comunicación, procesamiento y presentación, en función de los usuarios finales.

Ponniah la define como la estructura que unifica los componentes del *DataWareHouse*, donde provee un marco general para su desarrollo y despliegue. Además plantea que define los estándares, mediciones, diseño general y técnicas de soporte. (Ponniah, 2001).

En la figura 3 se presenta la arquitectura de la solución:

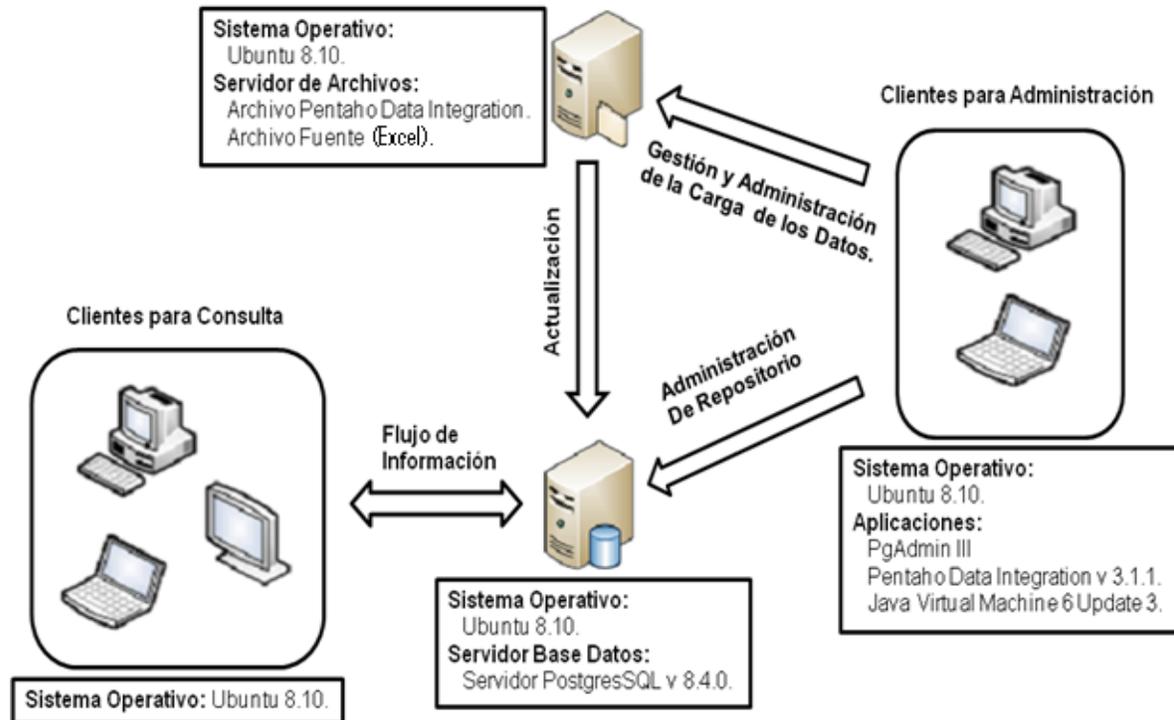


Figura 3: Arquitectura de la Solución.

La arquitectura definida se divide en 3 escenarios: El repositorio central, el de carga de datos y el componente de presentación. Como vía para el soporte del proceso se definieron un conjunto de herramientas para cada uno de estos escenarios.

Primeramente el repositorio central es el componente más importante sobre el cual se basa el sistema, la estructura de este se compone por un Gestor de Base de Datos PostgreSQL 8.4 donde se va a desplegar el sistema y como Sistema Operativo Ubuntu 8.10. Otro componente de igual importancia es el servidor de ficheros que almacenará las estructuras diseñadas en el Pentaho Data Integration v3.1.1 y las fuentes de los Excel que se deseen integrar al *DataWareHouse* permitiendo además su sincronización con los datos históricos. Las estaciones para consulta y administración se soportarán sobre Sistema Operativo Ubuntu 8.10 o Windows XP, además la estación de administración contará entre sus aplicaciones con el PgAdminIII como cliente de administración del Postgres, el Pentaho Data Integration para los procesos de ETL que se realizarán sobre el servidor de ficheros.

Las estructuras diseñadas que se realizan en el servidor de ficheros a partir del proceso ETL están

definidas en dos niveles de agrupación y concentrada en una misma instancia de la base de datos. En un nivel se encuentran las agregaciones diseñadas en función de los reportes más comunes y el otro están las estructuras con los datos detallados relacionados con las dimensiones propuestas.

Además se encuentran dos bases de datos más dentro del servidor, una llamada metadatos donde se guardarán todos los metadatos de las estructuras dimensionales, la cual es gestionada por la herramienta utilizada para los procesos de ETL y la llamada intermedia que es la encargada de almacenar la información de los procesos de ETL, donde aparecerán las estructuras necesarias para realizar la carga de los datos hacia el Almacén de Datos.

2.4 Conclusiones del Capítulo

En este capítulo se realizó un análisis que permitió describir la solución que se desarrolla, así como se explicó el proceso del negocio realizado para un mejor entendimiento de los usuarios que no se encuentran dentro de este y así comprendan todo el proceso realizado. Se definió como tema de análisis a el Control del Consumo de Portadores Energéticos en las oficinas de la ONE, obteniendo con esto las áreas sobre la cual se realizarán análisis de la información, se establecieron los roles y los permisos para lograr una buena seguridad de la BD. Se determinaron las reglas del negocio a medir, para lograr una estandarización del envío de la información, quedaron definidos los requisitos funcionales, no funcionales, de información, multidimensionales, para dar respuesta a las necesidades de los usuarios y conocer las condiciones, propiedades que tanto el sistema como el producto deben cumplir, se especificaron y detallaron los casos de uso del sistema, para lograr un mejor entendimiento de las necesidades de los usuarios como de las funcionalidades del sistema. Fueron modelados los diseños lógico, definiéndose en los mismo 6 dimensiones y 3 tablas de hechos, así como sus medidas y las políticas de seguridad a seguir en el sistema, finalmente se muestra la arquitectura de la solución para lograr una mejor comprensión de cómo funciona el sistema en su totalidad.

Capítulo III: Implementación y Pruebas

Introducción

En el desarrollo de este capítulo se abordará todo lo referente a la implementación en donde se reflejan la construcción del modelo de datos físico así como la estructura que poseen los datos del mismo, los usuarios y privilegios que se van a definir y la guía de implantación del sistema. Se describe además todo el proceso de carga de los datos al almacén de datos y se aplican de una manera clara las pruebas de rendimiento y estrés, además de las pruebas de Carga y Volumen, se procedió a validar la solución propuesta a través del empleo de las listas de chequeo y la carta de aceptación del cliente.

3.1 Modelos de Datos Físico

Un modelo de datos es un tipo de lenguaje que se orienta principalmente a describir una Base de Datos. Estos se pueden clasificar dependiendo de los tipos de conceptos que ofrecen para describir la estructura de la base de datos. En la actualidad existen los modelos de datos conceptuales o de alto nivel, los cuales están orientados a la descripción de estructuras de datos y restricciones de integridad, además disponen de conceptos muy cercanos al modo en que la mayoría de los usuarios percibe los datos; existen los Modelos de Datos Lógicos que son orientados a las operaciones más que a la descripción de una realidad y existen igualmente los modelos de datos físicos, (ver Anexo 2), que no son más que estructuras de datos a bajo nivel implementadas dentro del propio manejador y proveen conceptos que describen los detalles de cómo se almacenan los datos en el ordenador.

Es por esto que los conceptos de los modelos físicos están dirigidos al personal informático y no a los usuarios finales. Entre estos dos extremos se encuentran los modelos lógicos, cuyos conceptos pueden ser entendidos por los usuarios finales, aunque no están demasiado alejados de la forma en que los datos se organizan físicamente. (MASSACHUSETTS INSTITUTE, 2008).

3.1.1 Estructuras de Datos

Las estructuras de datos son una colección de datos cuya organización se caracteriza por las funciones

de acceso que se usan para almacenar y acceder a elementos individuales de datos; se pueden definir también como una forma de organizar un conjunto de datos elementales con el objetivo de facilitar su manipulación. Por debajo de estas estructuras hay un nivel, en el cual se sitúan los archivos, discos, particiones y espacios de tablas. La buena práctica con dichos elementos y el dominio pleno de los mismos inciden grandemente en el éxito de la solución. (EQUIPO06BD, 2008)

Esquemas y Tablas:

En una Base de datos Relacional, el Esquema define sus tablas, sus campos en cada tabla y las relaciones entre cada campo y cada tabla. El esquema es generalmente almacenado en un Diccionario de Datos. Aunque generalmente el esquema es definido en un lenguaje de Base de datos, el término se usa a menudo para referirse a una representación gráfica de la estructura de base de datos. Este esquema se especifica durante el diseño, y no es de esperar que se modifique a menudo, aunque los datos que se almacenan en la base de datos pueden cambiar con mucha frecuencia: ya que se insertan, se actualizan y se modifican. (EQUIPO06BD, 2008). **Los esquemas que fueron creados se describen a continuación:**

- **Esquema Electricidad:** Contiene a la tabla hecho de electricidad el cual controla todo lo referente al consumo de electricidad.
- **Esquema Combustible:** Contiene a la tabla hecho de combustible el cual controla todo lo referente al consumo de combustible.
- **Esquema General:** Contiene a la tabla hecho portadores energéticos el cual controla todo lo referente al consumo de portadores energéticos.
- **Esquema Dimensiones:** Contiene agrupadas a todas las tablas dimensiones que almacenan la información necesaria para realizar los análisis identificados para el control del consumo de los portadores energéticos.

Para más información sobre estas tablas las cuales fueron generadas a partir de las dimensiones y hechos que fueron generados en el análisis y diseño de la solución. (Consultar el Epígrafe 2.4.2).

Las tablas por su parte se refieren al tipo de modelado de datos, donde se guardan los datos recogidos por un programa. Su estructura general se asemeja a la vista general de un programa de Hoja de cálculo, (Ver Anexo 3). (EQUIPO06BD, 2008)

Restricciones y Secuencias:

Restricciones: Son condiciones que obligan el cumplimiento de ciertas reglas en la base de datos. Algunas no son determinadas por los usuarios, ya que son definidas por el mismo gestor de BD. Otras restricciones si las puede definir el usuario, por ejemplo, usar un campo con valores enteros entre 1 y 10. Las restricciones proveen un método de implementar reglas en la base de datos, así como restringen los datos que pueden ser almacenados en las tablas. La secuencia de las llaves debe ser incrementada según se definan siempre empezando por un valor mínimo hasta un máximo.

Secuencias: Son atributos que se van a ir incrementando secuencialmente durante la entrada de datos a la BD, un ejemplo de estos tipos de atributos son las llaves primarias:

Tabla de atributos auto-incrementales o secuenciales:

Secuencia	Esquemas	Propietario	Incrementad o	Valor Mínimo	Valor Máximo
dimactividad_dim_actividad_seq	dimensiones	Especialistas	1	1	2147483647
dimcombustible_dim_combustible_seq	dimensiones	Especialistas	1	1	2147483647
dimelectricidad_dim_electricidad_seq	dimensiones	Especialistas	1	1	2147483647
dimoficina_dim_oficina_seq	dimensiones	Especialistas	1	1	2147483647
dimtemporal_dim_temporal_seq	dimensiones	Especialistas	1	1	2147483647
dimvehiculo_dim_vehiculo_seq	dimensiones	Especialistas	1	1	2147483647

Tabla 2: Tabla Atributos Secuenciales.

Índices:

Las políticas de indexado son restricciones que el programador de BD define a la hora de indexar o crear un índice. Un índice es una estructura de disco asociada con una tabla o una vista que acelera la recuperación de filas de la tabla o de la vista. Contiene claves generadas a partir de una o varias columnas de la tabla o la vista. Dichas claves están almacenadas en una estructura (árbol b) que permite que SQL Server busque de forma rápida y eficiente la fila o filas asociadas a los valores de cada clave. (microsoft, 2010).

Una tabla o una vista pueden contener los siguientes tipos de índices:

- **Agrupado:**

- Los índices agrupados ordenan y almacenan las filas de los datos de la tabla o vista de acuerdo con los valores de la clave del índice. Son columnas incluidas en la definición del índice. Sólo puede haber un índice clúster por cada tabla, porque las filas de datos sólo pueden estar ordenadas de una forma.

- **No agrupado:**

- Los índices no agrupados tienen una estructura separada de las filas de datos. Contiene los valores de clave de índice no agrupado y cada entrada de valor de clave tiene un puntero a la fila de datos que contiene el valor clave.

A continuación se mostrarán aquellos índices que genera automáticamente el gestor, ejemplo a la hora de crear una llave primaria, cuando se crea un tabla o cuando se identifica una columna según la llave primaria coincidiendo con las propiedades o características que poseen los índices agrupados. Es significativo resaltar que los usuarios también pueden definir índices según las necesidades específicas de cada uno de estos.

Tabla de Índices:

Índice	Tablas	Esquemas	Tipo	Campo	Único	PK
PK4	dim_actividad	dimensiones	btree	dim_actividad	X	X
PK6	dim_combustible	dimensiones	btree	dim_combustible	X	X
PK6_1	dim_electricidad	dimensiones	btree	dim_electricidad	X	X
PK2	dim_oficina	dimensiones	btree	dim_oficina	X	X
PK3	dim_temporal	dimensiones	btree	dim_temporal	X	X
PK5	dim_vehiculo	dimensiones	btree	dim_vehiculo	X	X
PK1	hecho_consumo_combustible	combustible	btree	dim_oficina, dim_actividad, dim_combustible,	X	X

				dim_temporal, dim_vehiculo		
PK1_1	hech_consumo_electricidad	electricidad	btree	dim_electricidad, dim_temporal, dim_oficina	x	x
PK10	hech_portadores_energeticos	general	btree	dim_oficina, dim_temporal	x	x

Tabla 3: Tabla Índices.

Describir Artefactos Definición de Lenguaje de Datos (DDL): El DDL es el principal artefacto de la construcción en CADM (esquema de datos o modelo del servidor) lo constituyen los scripts de creación de los objetos de la base de datos. Estos scripts se construyen a partir de la transformación del esquema de datos. (Quintero, 2008). Este artefacto es un script que forma parte del código fuente del expediente de proyecto.

3.1.2 Procesos Extracción Transformación y Carga (ETL)

El proceso de ETL se basa en controlar la fuente, la transformación correspondiente y el destino de los datos en todo el proceso. La transformación de los datos se hará de acuerdo a las reglas que se definieron en el negocio. Se definen transformaciones tales como: cambios de formato que aseguran la unicidad y estandarización de los tipos de datos, por ejemplo la fecha, la cual deberá ser tratada como un solo tipo de dato y siempre de la misma forma.

Estos procesos ETL en este sistema en particular no va a tener la amplitud que normalmente estos requieren debido a que la información histórica se encuentra con un alto grado de limpieza, estandarización y calidad. Esto ocurre como consecuencia de los pasos previos que se realizan antes de llenar los Excel. Sólo habría que centrar los esfuerzos en dos áreas específicas: la extracción de los datos de las fuentes y la carga hacia el Almacén de Datos.

Para estos procesos ETL se utilizarán las ventajas que brinda la herramienta Pentaho Data Integration, como se explicó en el Capítulo 1. Con este fin el repositorio fue preparado con un conjunto de tablas que auxiliarían las estrategias definidas para la carga de nomencladores.

Carga de Nomencladores: El momento en el que los datos que provienen de la fuente son incluidos en el

CAPÍTULO III: IMPLMETACIÓN Y PRUEBAS

sistema destino habiendo pasado por un proceso de verificación recibe el nombre de fase de carga. Esta fase depende de los requerimientos de la organización y el proceso abarca una gran gama de acciones dado que por ejemplo algunas bases de datos sobrescriben los datos antiguos con la información nueva que reciben. Por su parte los almacenes de datos mantienen un historial de los registros permitiendo con esto que puedan efectuarse auditorías y que se disponga de trazas del comportamiento de un determinado valor a lo largo del tiempo. (Kimball y otros).

A continuación se muestran una lista de nomencladores que forman parte de la solución:

Tabla de Nomencladores:

Nomencladores	Valores
Nombre de Oficina	ONE sede central, ONE de Marianao, ONE del Vedado, etc.
Nivel de Oficina	Nacional, Provincial, Municipal.
Tipo de Combustible	Gasolina, Diesel, Grasas, Lubricantes.
Tipo de Actividad	Servicio, Administrativa.
Marca de Vehículo	Lada, Niva, Cherry, Fiat, Tollota, etc.
Tipo de Vehículo	Auto, Camioneta, Panel, Motor, Camión, Tractor, Ómnibus, etc.
Modelo de Vehículo	Son todo tipo de modelo que los vehículo pueden tener, ejemplo: jumpy, berlingo, etc.

Tabla 4: Tabla Nomencladores.

En la Tabla 5 se muestra el proceso de mapeo de los datos desde las fuentes los Excel, hasta el destino el Almacén de Datos.

Destino			Fuente			
Nombre de la tabla	Nombre de la columna	Tipo de dato	Nombre de la fuente de datos	Nombre de la tabla	Nombre de la columna	Tipo de dato
hech_consumo_combustible	consumo_real	integer	Excel Combustible	Consumo Portadores Energéticos	real(Tm)	interger
hech_consumo_combustible	consumo_asignado	integer	Excel Combustible	Consumo Portadores Energéticos	asignado(Tm)	interger
hech_consumo_combustible	cant_vehiculos	integer	Excel Combustible	Consumo Portadores Energéticos	cantidad Vehículos	interger
hecho_consumo_electricidad	consumo_real	integer	Excel Electricidad	Consumo Portadores Energéticos	Real(Tm)	interger

CAPÍTULO III: IMPLMETACIÓN Y PRUEBAS

hecho_consumo_el ectricidad	consumo_asignado	integer	Excel Electricidad	Consumo Portadores Energéticos	Asignados(Tm)	interger
dim_actividad	tipo	varchar	Excel Combustible	Actividades	Tipo	varchar
dim_actividad	descripcion	varchar	Excel Combustible	Actividades	Descripción	varchar
dim_combustible	tipo	varchar	Excel Combustible	Combustibles	Tipo	varchar
dim_combustible	descrpcion	varchar	Excel Combustible	Combustibles	Descripción	varchar
dim_oficina	nombre	varchar	Excel Combustible	Oficinas	Nombre	varchar
dim_oficina	codigo	varchar	Excel Combustible	Oficinas	Código	varchar
dim_oficina	nivel	varchar	Excel Combustible	Oficinas	Nivel	varchar
dim_oficina	direccion	varchar	Excel Combustible	Oficinas	Dirección	varchar
dim_vehiculo	marca	varchar	Excel Combustible	Vehículos	Marca	varchar
dim_vehiculo	modelo	varchar	Excel Combustible	Vehículos	Modelo	varchar
dim_vehiculo	tipo	varchar	Excel Combustible	Vehículos	Tipo	varchar
dim_vehiculo	indice	integer	Excel Combustible	Vehículos	Índice	integer
dim_vehiculo	indicee	varchar	Excel Combustible	Vehículos	Índice	varchar
dim_electricidad	descripcion	varchar	Excel Electricidad	Electricidad	Descripción	varchar

Tabla 5: Mapa Lógico de los Datos.

Describir Artefacto Lenguaje de Manipulación de Datos (DML): Este artefacto es un script el cual debe ser capaz de generar los nomencladores existentes en la solución al ser cargado en la BD, este script forma parte del código fuente del expediente de proyecto y de las herramientas de acceso a los SGBD, las cuales sirven para escribir e interpretar instrucciones SQL o DDL, que se envían al motor de bases de datos para ser ejecutadas. (Quintero, 2008).

3.1.3 Usuarios y Privilegios

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

Usuarios y Roles:

Los usuarios y roles definidos en la Base de Datos tributan a garantizar la seguridad de la misma. Se

establecieron los siguientes:

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

Privilegios:

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

- **Administrador:** Este usuario se le asigna los derechos de *Owner, Select, Update, Insert, Delete, Refresh y Trigger* sobre la estructura de la base de datos.
- **Analista:** Al contrario este tipo de usuario solo tiene derechos de *Select* los datos almacenados en el Almacén.
- **Programador ETL:** Por el rol que este usuario desempeña a hora de interactuar con la BD, se le es asignado los privilegios de *Select, Insert, Update, Delete, Refresh y Trigger* de los datos almacenados en el Almacén.

El artefacto que genera dichas acciones es el Lenguaje de Control de Datos (*DCL*), que se encarga de reflejar, todo el proceso de acceso a la base de datos, es un lenguaje utilizado en el control de la información, donde los usuarios solo puedan acceder a los datos que les fueron concedidos privilegios a él. Este artefacto es un script que forma parte del código fuente del expediente de proyecto.

Conociendo los detalles más fundamentales del modelo de dato físico, se procederá a la explicación paso a paso para la implantación del sistema.

3.2 Guía de Implantación

Las guías de implantación contienen los pasos necesarios para la implantación de cualquier sistema informático, no es más que las necesidades que presenta el software para su instalación. Antes de abordar los pasos necesarios para lograr esta implantación, es preciso conocer los requerimientos del sistema.

3.2.1 Requerimientos

Estos requerimientos fueron definidos en el Epígrafe 2.3.9 los cuales constituyen requisitos no funcionales que el sistema debe cumplir, los más importantes que se abordan en ese epígrafe son los de interfaces de hardware y software. Luego de conocer estas necesidades, se describen a continuación los pasos de implantación de la solución.

3.2.2 Secuencia de Pasos

Para la implantación de la Base de Datos, se deben seguir una serie de pasos que a continuación se detallan:

- Debe estar instalado el gestor de base de datos PostgreSQL 8.4.
- Debe estar instalada una herramienta de administración de base de datos, la cual ya fue definida, el PgAdmin III.
- Se debe crear una base de datos nueva utilizando la herramienta de administración de bases de datos deseada.
- Luego de tener la BD nueva lista, se carga el script DDL para crear la estructura física del Almacén.
- Teniendo la estructura de la BD están creadas las condiciones para correr el script lenguaje de manipulación de datos DML, quedando así cargados los nomencladores en el Almacén.
- Al estar la BD creada estructuralmente con los nomencladores cargados se debe correr el script lenguaje de control de datos para los usuarios y permisos definidos en el Almacén.
- Para poder acceder a la BD desde otra terminal de trabajo, se debe modificar los archivos de configuración de Postgres.
- Instalar la maquina virtual de java 6 Update 3.
- Instalar el servidor Apache Tomcat 5.5 donde correrá la aplicación Mondria.
- Luego de tener configurado el Postgres y tener la BD cargada, se procede a copiar la aplicación Mondria en la carpeta raíz del Tomcat.
- Finalmente se debe configurar el Apache Tomcat y los archivos del Mondria que permitirá la conexión a la BD.

Culminado este proceso, el almacén cuenta con todos los datos, tanto de las tablas de dimensiones como

de hechos, ya integrados y almacenados de una manera consistente. En este punto las condiciones están creadas para comenzar el desarrollo de la capa de visualización que permitirá mostrar al usuario los principales reportes que responden a sus necesidades de información en función de los datos almacenados.

3.4 Procesos de Inteligencia del Negocio (BI)

Al igual que otros conceptos o términos, el de Inteligencia de Negocios no escapa a la variedad de interpretaciones. Sin embargo queda esencialmente claro que: “no es una metodología, software, sistema o herramienta específica, es más bien una colección de tecnologías que van desde arquitecturas para almacenar datos, metodologías, técnicas para analizar información y software, entre otros, con un fin común para el apoyo a la toma de decisiones”. (Caramazana, 2005).

Después de analizar el concepto anterior y de tener un conocimiento básico sobre este tema se puede afirmar que la Inteligencia de Negocio es considerada un término "agrupador". El hecho de que sea considerado como una colección de conceptos le da un poder enorme, pues pueden integrarse funciones que tradicionalmente estaban separadas, tales como el acceso de datos, reporte, explotación, pronóstico y análisis.

El desarrollo de un almacén de datos, como solución de la investigación, permitirá almacenar la información concerniente a la energía eléctrica y combustible consumido en las oficinas de la ONE para transformarlos en datos útiles a las personas encargadas de tomar decisiones, para la definición de estrategias que permitan disminuir el consumo de electricidad y combustible de estas oficinas. Partiendo de esto se especifica el tema de análisis identificado, así como los libros de trabajo en las que se divide este tema, que no son más que las áreas de análisis las cuales agrupar los reportes construidos. Cada uno de estos componentes se describe en el Artefacto Arquitectura de Información (AI).

Además de la AI otro de los elementos fundamentales donde se organiza la información estructuralmente son los cubos OLAP los cuales pueden poseer más de tres dimensiones por lo que son llamados hipercubos. Los cubos OLAP presentan diferentes métodos de creación, en el caso concreto de la solución se desarrollaron 3 cubos multidimensionales principales, Combustible, Electricidad y Portadores, apoyándose en la herramienta Pentaho Workbench la cual en el capítulo 1 fueron abordados las

facilidades y diferentes características que este producto facilita a la hora de crear estos cubos. A continuación se muestran el diseño estos cubos multidimensionales.

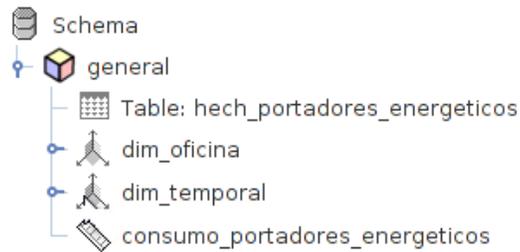


Figura 4: Cubo general.

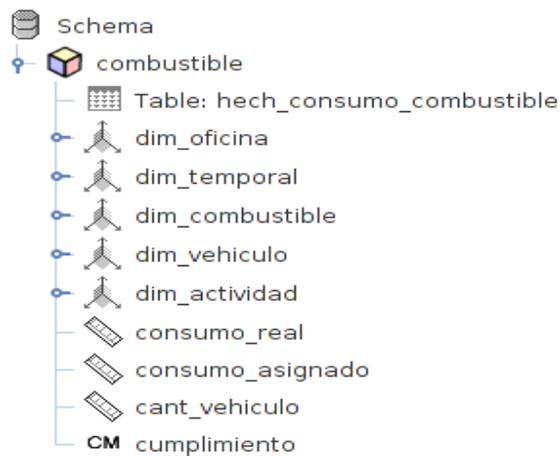


Figura 5: Cubo combustible.

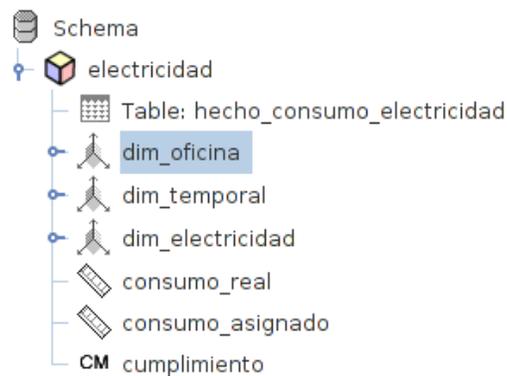


Figura 6: Cubo electricidad.

Como bien se observa en la figuras anteriores, estos cubos OLAP son definidos por cada tabla hecho, debido a que estas tablas son las fuentes de información principal de la organización estructural de los datos, además se observan las tablas dimensiones relacionadas con este hecho así como las medidas del mismo, en el Epígrafe 2.2.2 se abordan con más detalles estas tablas.

Luego de completar la solución, es de vital importancia garantizar su calidad y adherencia a los requisitos del cliente, esto se logra a través de la realización de validaciones y pruebas, aplicando esto se obtendrá un software que estará a la altura de un producto de alto nivel y con la calidad requerida.

3.5 Validación y pruebas

Los procesos de validación y pruebas son actividades que garantizan en gran medida que el software funcione adecuadamente y que además esté presente en él la calidad requerida. Existen varios tipos de métodos para validar el buen funcionamiento de los almacenes de datos, algunos de estos son conocidos como pruebas de rendimiento y estrés, de implantación, y de integración.

3.5.1 Listas de Chequeo Análisis

Las listas de chequeo se realizan con el objetivo de evaluar la calidad de los artefactos que se generan en el análisis de la solución. Mediante estas listas se pueden cubrir todos los aspectos aplicables en los temas de análisis, particularmente para el tema de consumo de portadores energéticos en las oficinas de la ONE. A continuación se muestran las listas que son aplicadas en el análisis las cuales forman parte del código del Expediente de Proyecto.

- Lista de Chequeo Especificación de Requisitos.
- Lista de Chequeo Especificación de las áreas de la organización.
- Lista de Chequeo Herramienta para la recolección y análisis de la información.

Al ser aplicados estos chequeos a los artefactos que se generaron en la fase de análisis, se detectaron 3 no conformidades de 62 chequeos realizados, estos errores detectados son básicamente errores semánticos. Con este resultado se observa la solidez del análisis realizado.

3.5.2 Listas de Chequeo de Diseño

La lista de chequeo de diseño es esencial para lograr un buen diseño de la solución. Es específica para cada sistema y pretende cubrir todos los aspectos aplicables en los temas de diseño e implementación, particularmente para los portadores energéticos. Estas listas son generadas y aplicables en la fase de diseño, a continuación se nombra este tipo de lista, la cual forma parte del código fuente del expediente de proyecto.

- Lista de Chequeo Modelo de Datos.

En esta lista de chequeo se puede encontrar una serie de estándares definidos con el objetivo de logra la realización de un diseño estandarizado, definiendo pautas claves para la conformación de una estructura concisa y entendible.

3.5.3 Validación de requisitos por el cliente

En la validación del almacén de datos para verificar la aceptación de este producto, estuvo presente la especialista de la ONE:

- Elena Leonila Fernández García. Representante de la ONE en la UCI.

La compañera estuvo de acuerdo con el resultado obtenido en la solución que se documenta. Durante el proceso de análisis y diseño estuvo supervisando cada avance, quedando satisfecha con el diseño lógico y la arquitectura de la solución que se generó al finalizar esta fase. Es importante resaltar que durante la implementación de la solución se obtuvieron resultados tangibles, como el modelo físico de la BD y la aplicación Mondrian para la visualización de información del Almacén de Datos para el Control del Consumo de Portadores Energéticos, con los que la especialista de la ONE estuvo satisfecha.

3.5.4 Pruebas de Rendimiento y Estrés

Las Pruebas de Rendimiento se ejecutan tanto para determinar cómo responde un sistema ante una cierta carga, como para validar otros atributos relacionados con la calidad, como pueden ser la escalabilidad o el uso de recursos entre otros. (HeadQuarter, 2010).

La arquitectura general que se utilizará para la realización de las pruebas serán 3 estaciones clientes con

el Jmeter configurado directamente con el servidor de BD, herramienta utilizada para realizar este tipo de pruebas. Dos de las estaciones clientes sólo limitarán su uso a realizar peticiones indefinidamente al servidor y la otra para llevar las estadísticas con el número de muestras definido en 50. Se le realizarán pruebas con cantidades diferentes de usuarios concurrentes, 5 y 10 respectivamente, para realizar el análisis de los resultados debido a que según los especialistas de la UCI nunca existirán más de 25 usuarios registrados en el servidor y en general la concurrencia será mínima. Las consultas se realizarán sobre las agregaciones definidas para este fin. Se considera necesario aclarar que el servidor utilizado para las pruebas no posee todas las prestaciones de un servidor profesional debido a que se utilizó una estación cliente con características mejoradas. Esto afecta la calidad de las pruebas pero su objetivo es dar una idea del rendimiento de la solución. Ver Figura 7.

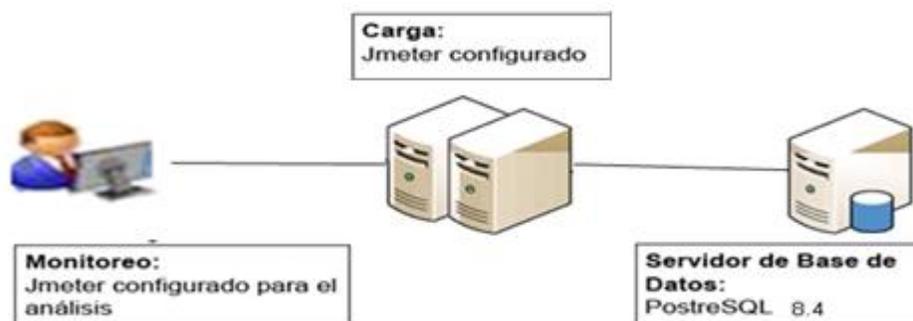


Figura 7: Arquitectura para la realización de la prueba.

Al aplicar este tipo de prueba se obtuvieron los siguientes resultados:

Prueba No 1: Comportamiento del consumo real de combustible de las oficinas.

- Tablas involucradas: hech_consumo_combustible, dim_oficina.
- Consulta realizada: `SELECT combustible.hech_consumo_combustible.consumo_real FROM combustible.hech_consumo_combustible, dimensiones.dim_oficina WHERE combustible.hech_consumo_combustible.dim_oficina = dimensiones.dim_oficina.dim_oficina.`
- Cantidad de usuarios:
 - 5 usuarios concurrentes.

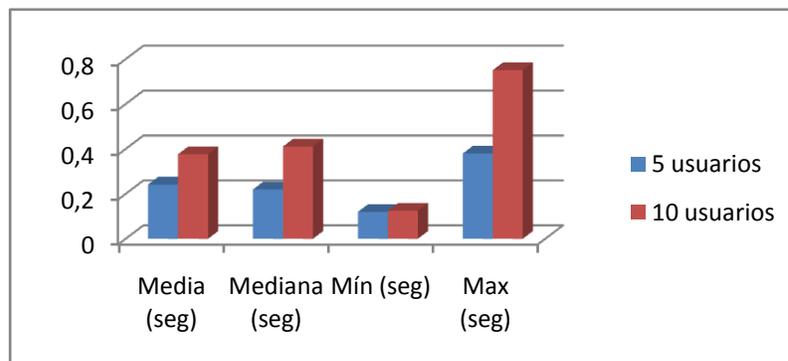
	Media (seg)	Mediana (seg)	Mín (seg)	Max (seg)

Resultados	0,242	0,220	0,120	0,380
------------	-------	-------	-------	-------

- 10 usuarios concurrentes.

	Media (seg)	Mediana (seg)	Mín (seg)	Max (seg)
Resultados	0,377	0,411	0,125	0,751

- Gráfico de relación entre las pruebas.



Como puede apreciarse los resultados se enmarcan en 4 variables, aportadas por la herramienta Jmeter, de significativo valor para la presentación de los resultados de las pruebas: la media, valor de la suma aritmética de los tiempos de respuesta dividido entre 2, la mediana, valor de la variable que deja el mismo número de datos antes y después que él, una vez ordenados estos, de acuerdo con esta definición el conjunto de datos menores o iguales que la mediana representarán el 50% de los datos y los que sean mayores que la mediana representarán el otro 50% del total de datos de la muestra; el valor mínimo, que se refiere al tiempo de respuesta menor de todos los usuarios que hicieron peticiones concurrentes y el valor máximo que, similarmente, es el tiempo mayor de respuesta a todos los usuarios.

En las gráficas anteriores los tiempos de respuestas oscilaron en dependencia de la cantidad de filas que se recuperen en la consulta. En general los resultados obtenidos son satisfactorios evidenciándose un aumento casi simétrico entre cada una de las configuraciones realizadas. Destacándose los mayores

tiempos en la agregación consumo mensual por oficina debido a que la consulta devuelve más cantidad de tuplas a cada usuario, convirtiéndose en un proceso un poco más lento.

3.5.5 Pruebas de Volumen y Carga

Las pruebas de volumen son pruebas típicas de entornos que utilicen bases de datos. Las mismas se realizan con el objetivo de analizar el comportamiento del sistema o base de datos con volúmenes de datos almacenados lo más similar posible a los esperados en la explotación real del sistema.

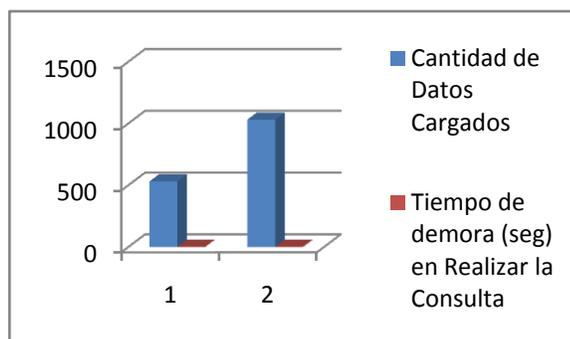
Una prueba de carga se realiza para analizar el comportamiento de una aplicación ante una carga determinada. Esta carga puede un número de usuarios esperados, ejecutando un número de transacciones durante un tiempo determinado. Otro objetivo principal de esta prueba es buscar consultas mal diseñadas, consultas candidatas a optimización, la necesidad de índices adicionales, código mal diseñado, tiempo de demora de respuesta de magnitudes inaceptables, hardware insuficiente, problemas de control de concurrencia, etc. El resultado de esta prueba nos dará el tiempo de respuesta de todas las transacciones críticas.

Para realizar estos tipos de pruebas se utilizarán los beneficios que brinda la herramienta *PgDataGenerator*. Este es un generador de carga diseñado para la realización de pruebas de este tipo. Genera carga por diversos protocolos, ya sea, HTTP, SQL, etc. Realiza carga variable, en niveles de concurrencia, número de veces, tiempo. Al aplicarse este tipo de prueba sobre el almacén se observó que:

Al realizarse a la tabla hecho_consumo_electricidad la siguiente consulta, `select * FROM electricidad.hecho_consumo_electricidad`, se obtuvo:

Cantidad de Datos Cargados	530	1030
Tiempo de demora (seg) en Realizar la Consulta	0,091	0,165

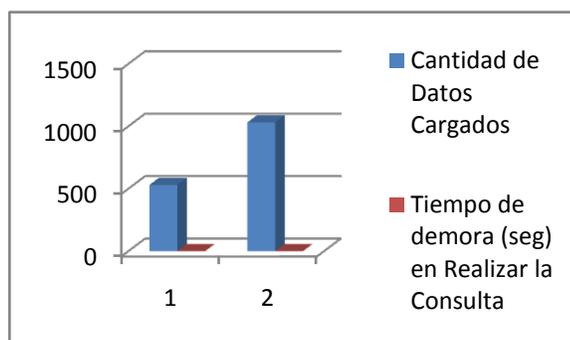
En la siguiente gráfica se muestra el comportamiento de este resultado:



Que al realizarse a la tabla hecho_consumo_combustible la siguiente consulta, `select * FROM combustible.hech_consumo_combustible,` se obtuvo que:

Cantidad de Datos Cargados	530	1030
Tiempo de demora (seg) en Realizar la Consulta	0,9	0,16

En la siguiente gráfica se muestra el comportamiento de este resultado:



Al poblar el almacén no se presentó problemas de límite de capacidad, ni de volumen. Las llaves autogenerated no se salieron del rango especificado, ni se detectaron problemas con los tipos de datos definidos en el paso de diseño. Todo esto verifica que PostgreSQL como gestor utilizado y el diseño de las estructuras de la base de datos implementadas soportan completamente el almacenamiento de los niveles de información requeridos para la puesta en producción del Almacén.

Como pudo apreciarse los resultados de esta prueba se enmarcan en 2 variables básicas, la cantidad de datos cargados y el tiempo de respuesta de la BD ante una consulta realizada. Este proceso se compuso de dos fases, una la carga de la BD con cierta cantidad de datos y comparar el tiempo de respuesta con el tipo de la segunda fase que se cargaría la BD con una mayor cantidad de datos que en la primera fase. En

las tablas y gráficas anteriores se muestran estas comparaciones, notándose que la diferencia de los tiempos de respuesta de la BD para ambas fases es aceptable dado que a pesar de que la segunda fase posee el doble de datos de la primera su tiempo de respuesta no dobla al tiempo de respuesta de la primera fase.

3.5.6 Pruebas de Implantación

Para verificar la calidad del proceso de implementación se procedió a establecer un modelo de casos de prueba que facilitan la identificación de resultados verídicos y eficientes. A continuación se explican los tres tipos de escenarios de implantación que fueron aplicados.

Pre-Condiciones	Resultados Esperados	Pos-Condiciones
Antes de la implantación del sistema, se debe tener instalado el PostgreSQL y para su administración PgAdminIII. Además para la aplicación Web se debe tener corriendo el servidor apache Tomcat.	Al tener las pre-condiciones específicas se obtienen las condiciones necesarias para la implantación del sistema en su totalidad.	Se crea la Base de Datos donde se montará el Almacén y se copia el sitio en la raíz del servidor apache que permitirá mostrar gráficamente los datos almacenados.
Se debe tener creada la BD, para correr el script DDL.	Tener la estructura de la BD.	Se tienen las condiciones necesarias para proceder a la carga de los nomencladores.
Se debe correr el script DDL para la carga del DML.	Tener los nomencladores cargados en las diferentes estructuras de BD.	Se tiene el almacén creado listo para su uso y definirle los usuarios y permisos.
Se debe tener la BD creada, así como la estructura de la misma para la carga del DCL	Tener los usuarios y permisos cargados en el almacén.	Se tiene el Almacén listo para su uso.
Para visualizar la información del Almacén ya listo, se debe configurar los archivos que permiten conectar el sitio con la BD.	La visualización de los datos.	El sistema está preparado para su utilización.

Tabla 6: Pruebas de Implantación

Conclusiones del Capítulo

En este capítulo se abordaron los temas que dieron paso a la implementación e implantación del Almacén de Datos y cómo lograr que este producto quede con la calidad requerida. Se dio a conocer como quedó físicamente el Almacén logrando una estructura robusta entre las relaciones de las tablas multidimensionales, estructuras las cuales fueron agrupadas por esquemas para una mejor organización de la BD. Se definieron en estas estructuras índices con el objetivo de lograr rapidez y seguridad a la hora de realizar alguna consulta, el mantenimiento y seguridad de estas estructuras se logra a través de la definición de usuarios y permisos.

Se logró una buena sincronización entre la fuente de los datos y el destino al extraer y transformar la información a como se carga en la BD, además de pasos que se logran mediante los procesos de ETL. Al tener los datos cargados en el almacén se lograron transformar a estos en información, ésta en conocimientos, para realizar acciones y análisis sobre estos conocimientos, fueron transformados en planes. Los pasos mencionados anteriormente son realizados en el proceso de BI.

Finalmente se realizaron pruebas de estrés, de volumen y de implantación, esta última se refiere a los pasos necesarios a seguir para una buena instalación de la aplicación y un buen uso de esta, fueron definidas listas de chequeos donde se cubren todos los aspectos aplicables en el tema de análisis y aspectos necesarios para cumplir con un buen diseño, garantizando de esta manera la calidad de la solución y el cumplimiento de los requerimientos del cliente.

Conclusiones

Se arribaron a las siguientes conclusiones:

- Se utilizó la tecnología de Almacén de Datos que permitió integrar fuentes de datos muy diversas y quedando esta información a disposición de los usuarios para su consulta de una manera eficiente, apoyando el proceso de toma de decisiones.
- La propuesta de modelos de recogida de información, cumplió con lo requerido sobre los datos que se necesitan para el análisis del consumo de los portadores energéticos.
- La estructura lógica modelada en la fase de análisis y diseño, cumplió con lo requerido en el negocio.
- Las estructuras dimensionales implementadas para el control de información soportaron el proceso de toma de decisiones, creándose con esto las condiciones necesarias para integrar la información.
- El Gestor de Base de Datos PostgreSQL garantizó la administración de los volúmenes de información que este tipo de solución necesita.
- Al cargarse los datos en el Almacén se observó que la presentación visual realizada en el Mondrian estuvo acorde con los datos almacenados.
- Las pruebas realizadas permitieron validar la calidad del análisis, la adherencia de los requisitos, el rendimiento del almacén y el crecimiento del volumen de datos de la BD.

Recomendaciones

En el desarrollo de este trabajo fueron tomadas en consideración muchas ideas, pero dadas las necesidades de atender aspectos de mayor importancia no todas fueron contempladas. Por lo que se recomienda para el desarrollo futuro del sistema:

- Mantener actualizado el contenido del almacén, incorporando una mayor cantidad de datos para que se realicen análisis más profundo de los mismos y este cumpla con el propósito por el cual fue desarrollado.
- Que se implemente minerías de datos sobre los datos del Almacén. Debido a que este tipo de análisis prepara, sondea y explora los datos para sacar la información oculta en ellos.

Bibliografía

Altacracks. 2007. *www.altacracks.com*. [En línea] 2007. [Citado el: 26 de 04 de 2010.]
<http://www.altacracks.com/programas-gratis/dtm-data-generator>.

ArPUG. 2009. *ArPUG Grupo de Usuario PosgreSQL. ArPUG Grupo de Usuario PosgreSQL*. [En línea] 2009.
[Citado el: 20 de 03 de 2010.] <http://www.arpug.com.ar/trac/wiki/PgAdmin>.

Artículo 55, de la constitución de la República Bolivariana. 2008. *Derecho de la seguridad Ciudadana. Derecho de la seguridad Ciudadana*. [En línea] 05 de 02 de 2008. [Citado el: 10 de 03 de 2010.]
http://www.derechos.org/ve/publicaciones/infannual/2002_03/16seguridaCIUDADANA.pdf.

Bernabeu, R. D. Hefesto. 2007. *Metodología propia para la construcción de un Data Warehouse. Metodología propia para la construcción de un Data Warehouse*. [En línea] 2007. [Citado el: 10 de 03 de 2010.] Disponible en: [<http://www.dataprix.com/es/hefesto-metodologia-propia-para-la-construccion-un-data-warehouse>].

Caramazana, Alberto. 2005. *Tecnologías y Metodologías para la Construcción de Sistemas de Gestión del Conocimiento*. Madrid : s.n., 2005.

cecre. 2009. *sourcews. sourcews*. [En línea] 2009. [Citado el: 10 de 03 de 2010.]
<http://www.sourcews.es/miguel-sebastian-visita-centro-control>.

Company Headquarters. 2009. *visual paradigm. visual paradigm*. [En línea] 2009. [Citado el: 20 de 03 de 2010.]
<http://www.visual-paradigm.com/>.

Copyright. 2008. *Definición.de. Definición.de*. [En línea] 2008. [Citado el: 10 de 03 de 2010.]
<http://definicion.de/gas-natural/>.

Curto, J. CIF vs MD. 2008. *Dos enfoques clásicos en el diseño de la arquitectura de una Data Warehouse. Dos enfoques clásicos en el diseño de la arquitectura de una Data Warehouse*. [En línea] 2008. [Citado el: 10 de 03 de 2010.] Citado en:[<http://bi-businessintelligence.blogspot.com/2009/01/cif-vs-md-dos-enfoques-clasicos-en-el.html>].

Definición.de. 2008. *Definición.de. Definición.de*. [En línea] 2008. [Citado el: 10 de 03 de 2010.]
<http://definicion.de/gas-natural/>.

Doc. Pedro Yobanis Piñero, Osiel Hernandez Calvo y otros. 2009. *Metodología para Soluciones de Integración y Análisis de Datos VFinal*. Habana : s.n., 2009.

- Duque, Raúl González. 2008. **Mundo Geek**. *Mundo Geek*. [En línea] 2008. [Citado el: 23 de 02 de 2010.] <http://mundogeek.net/archivos/2004/08/26/modelo-de-datos/>.
- EQUIPO06BD. 2008. **Modelos de Bases de Datos**. *Modelos de Bases de Datos*. [En línea] 2008. [Citado el: 23 de 04 de 2010.] <http://modelobdpnfi.wordpress.com/2010/04/06/base-de-datos-distribuidas/>.
- Ferri, Ramírez, M.J.Q y Hernández, J.O. 2004. **Introducción a la Minería de Datos**. s.l. : Editorial Pearson, 2004.
- Gómez, Isabel Criado y López, Paloma Sánchez. **NorSistemas**. *NorSistemas*. [En línea] Empresa Consultora Externa. [Citado el: 29 de 03 de 2010.] <http://www.csae.map.es/csi/silice>.
- Group, Development. 1996-2010.. **PostgreSQL**. *PostgreSQL*. [En línea] PostgreSQL Global, 1996-2010. [Citado el: 18 de 02 de 2010.] <http://www.postgresql.org/&ei=zuiHS72hNcP08Qbb8JGYDw&sa=X&oi=translate&ct=result&resnum=1&ved=0CA0Q7gEwAA&prev=/search%3Fq%3Dpostgresql%26hl%3Des>.
- Grupo soluciones. 2009. **Rational Rose Enterprise** . *Rational Rose Enterprise* . [En línea] 2009. [Citado el: 20 de 03 de 2010.] <http://www.rational.com.ar/herramientas/roseenterprise.html>.
- HeadQuarter. 2010. **test house**. [En línea] 2010. [Citado el: 01 de 05 de 2010.] <http://es.testhouse.net/index.php/servicios/pruebas-no-funcionales/pruebas-de-rendimiento>.
- HispaNetwork Publicidad, y Servicios, S.L. 2003. **glosario.net**. *glosario.net*. [En línea] 2003. [Citado el: 10 de 03 de 2010.] <http://energia.glosario.net/terminos-petroleo/hidr%F3geno-1958.html>.
- Hobbs y otros, Lilian. 2005. **Oracle Database 10g Data Warehousing**. EUA : ELSEVIER Digital Press, 2005.
- Imhoff y otros, Claudia, Galemmo, Nicholas y Geiger, Jonathan G. 2003. **Mastering Data Warehouse Desing, Relational and Dimentional Techniques**. EUA : Wiley Publishing Inc, 2003, 2003.
- Inmon y otros, William H. 2005. **Building the Data Warehouse**. EUA : Wiley Publishing Inc, 2005.
- Kimball y otros, Ralph. **The Data Warehouse Lifecycle Toolkit**. EUA : Wiley Publishing Inc.
- Kimball y otros, Ralph y Ross, Margy. 2002. **The Data Warehouse Toolkit**. EUA : Wiley Publishing Inc, 2002.
- Linux, Telemática &. 2009. **linux-itt**. *linux-itt*. [En línea] 2009. [Citado el: 10 de 03 de 2010.] <http://www.linux-itt.com/2009/07/el-consumo-domestico-de-energia-se.html>.
- MASSACHUSETTS INSTITUTE. 2008. **BASES DE DATOS DISTRIBUIDASHOMOGENEAS**. 2008. 14.
- MIC. 2008. **atenas**. *atenas*. [En línea] 2008. [Citado el: 10 de 03 de 2010.] <http://www.atenas.cult.cu/?q=node/6415>.

- Michael Kornspan. 2010. **ERwin**. *ERwin*. [En línea] Microsoft, 2010. [Citado el: 06 de 02 de 2010.] <http://erwin.com/>.
- Microsoft Dynamics. 2008. **Aprende Dynamics**. *Aprende Dynamics*. [En línea] 2008. [Citado el: 24 de 03 de 2010.] http://www.aprendedynamics.com/conceptos_olap.html&usg=__rxFg8IP7iTgKZpMDm65p1gu8Rfc=&h=243&w=187&sz=6&hl=es&start=13&um=1&itbs.
- microsoft. 2010. **msdn**. [En línea] 2010. [Citado el: 23 de 04 de 2010.] <http://msdn.microsoft.com/es-es/default.aspx>.
- Microsoft. 2008. **SqlServer**. *SqlServer*. [En línea] Microsoft., 2008. [Citado el: 17 de 02 de 2010.] <http://www.microsoft.com/latam/sqlserver/microsoft-oracle.aspx>.
- minbas. 2008. **atenas**. *atenas*. [En línea] 2008. [Citado el: 10 de 03 de 2010.] <http://www.atenas.cult.cu/?q=node/6415>.
- Mining, Pentaho Data. 2008. **pentaho**. *pentaho*. [En línea] 2008. [Citado el: 10 de 03 de 2010.] http://www.pentaho.com/products/data_mining/.
- OPEP. 2008. **info-petroleo**. *info-petroleo*. [En línea] 2008. [Citado el: 10 de 03 de 2010.] <http://www.info-petroleo.com/>.
- Oracle. 2010. **Oracle**. *Oracle*. [En línea] 2010. [Citado el: 18 de 02 de 2010.] <http://www.oracle.com/subscribe/index.html&prev=/search%3Fq%3DOracle%26hl%3Des&rurl=translate.google.com.cu&usg=ALkJrhjxkLOVlt5MBNinQjSWvVV4qKfQqQ>.
- Organismo Internacional de Energía. 2005. **Indicadores energéticos**. *Indicadores energéticos*. [En línea] 2005. [Citado el: 20 de 03 de 2010.] http://www.un.org/esa/dsd/resources/res_pdfs/publications/sdt_ind/Pub1222s_web_S.pdf.
- Peñaloza, Lucía Victoria Hernández. 2008. **Tesis para logra el título de Magíster: Diseño y Construcción de un Data Mart para la mantención de Indicadores de Sostenibilidad de la Industria del Salmón**. Chile : s.n., 2008.
- Pentaho. 2009. **Pentaho Open Source Business Intelligence**. *Pentaho Open Source Business Intelligence*. [En línea] Kettle Project, 2009. [Citado el: 2010 de 03 de 15.] <http://kettle.pentaho.org/>.
- Ponniah, Paulraj. 2001. **Data Warehousing Fundamentals**. EUA : Wiley Publishing Inc, 2001.
- Quintero, Juan Bernardo. 2008. **DIRECTRICES PARA LA CONSTRUCCIÓN DE**. 2008. 14.
- Sierra, Julio Ernesto Ortiz. 2009. **Diseño e Implementación de un Mercado de Datos para la Oficina Nacional de**

- Estadísticas. *Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas*. Ciudad de la Habana : s.n., 2009.
- Smith, Armstrong. 2006. *Darlene and Michel Oracle Discoverer 10g Handbook*. San Francisco, California : The McGraw-Hill Companies, 2006. 516.
- Softqanetwork. 2005. *Softqanetwork.com*. [En línea] 2005. [Citado el: 25 de 04 de 2010.] <http://www.softqanetwork.com/2005/10/jmeter/>.
- SPC. 1998. *datalyzer*. *datalyzer*. [En línea] 1998. [Citado el: 10 de 03 de 2010.] <http://www.datalyzer.com/spanish/products/q98/>.
- Studio, Power. 2008. *circutor*. *circutor*. [En línea] 2008. [Citado el: 10 de 03 de 2010.] <http://www.circutor.es/novedades/powerstudio/powerstudio-sp.html>.
- Talend. 2009. *Talend Open Data Solutions* . *Talend Open Data Solutions*. [En línea] 2009. [Citado el: 2010 de 03 de 15.] <http://es.talend.com/index.php>.
- Unión de Asociaciones, y Entidades de Atención al D. 2007. *UNAD*. *UNAD*. [En línea] 2007. [Citado el: 10 de 03 de 2010.] http://calidad.unad.org/asesoramiento/definicion_de_indicadores.html.
- Zepeda Sánchez, L. Z. 2008. *Metodología para el Diseño Conceptual de Almacenes de Datos*. *Universidad Politécnica de Valencia*. Valencia. 2008. 211 Páginas.

Glosario de Términos

Hidrocarburos: Son compuestos orgánicos formados únicamente por átomos de carbono e hidrógeno.

Indexados: Refiere a la acción de registrar ordenadamente información para elaborar índices en una BD.

Normalización: Es el proceso de elaboración, aplicación y mejora de las normas que se aplican a distintas actividades científicas, industriales o económicas con el fin de ordenarlas y mejorarlas.

Escalabilidad: Capacidad de un sistema informático de cambiar su tamaño o configuración para adaptarse a las circunstancias cambiantes.

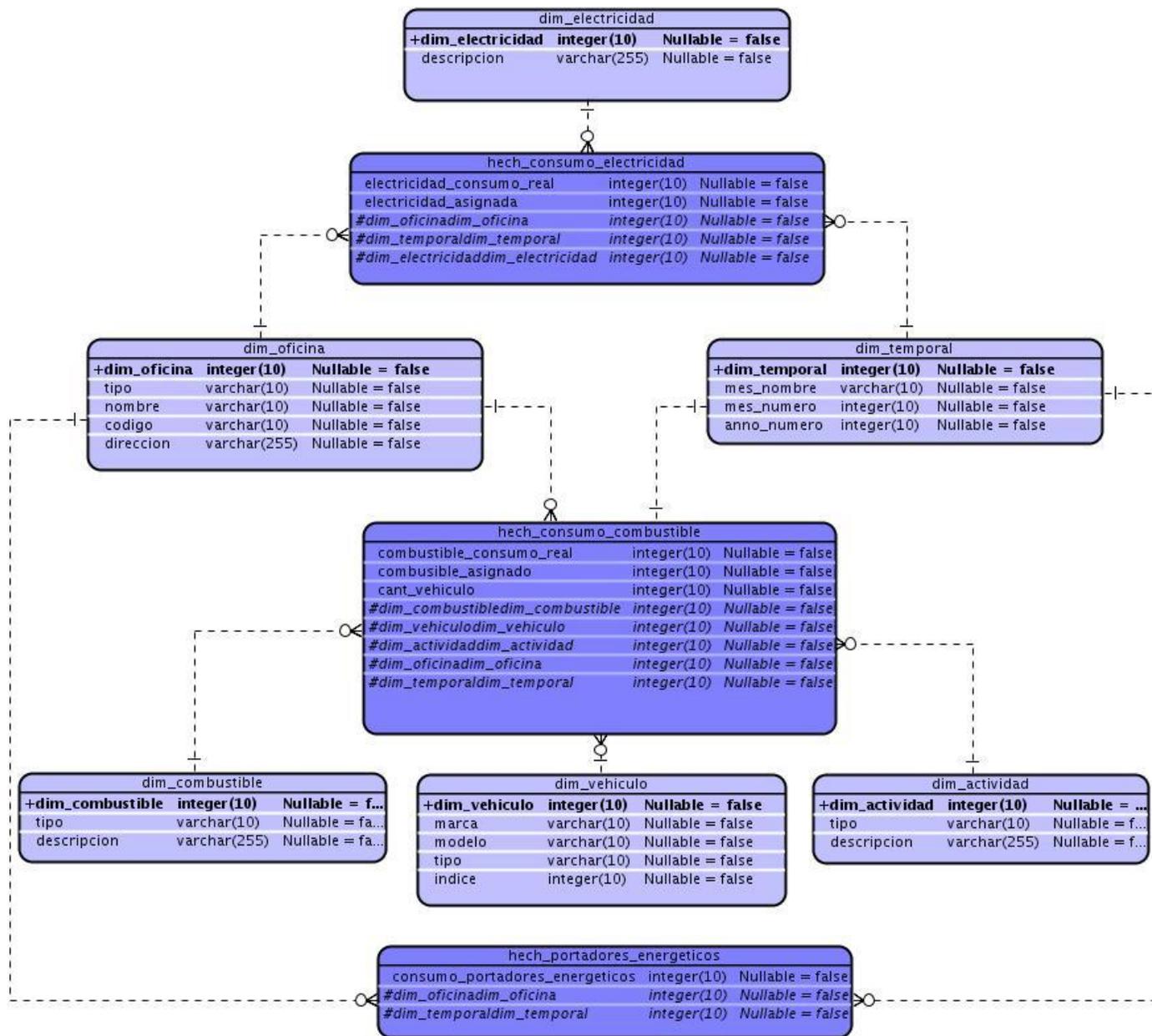
Plugins: Es un complemento para una aplicación que se relaciona con esta para aportarle una función nueva y generalmente muy específica.

Open source: Código abierto es el término con el que se conoce al software distribuido y desarrollado libremente.

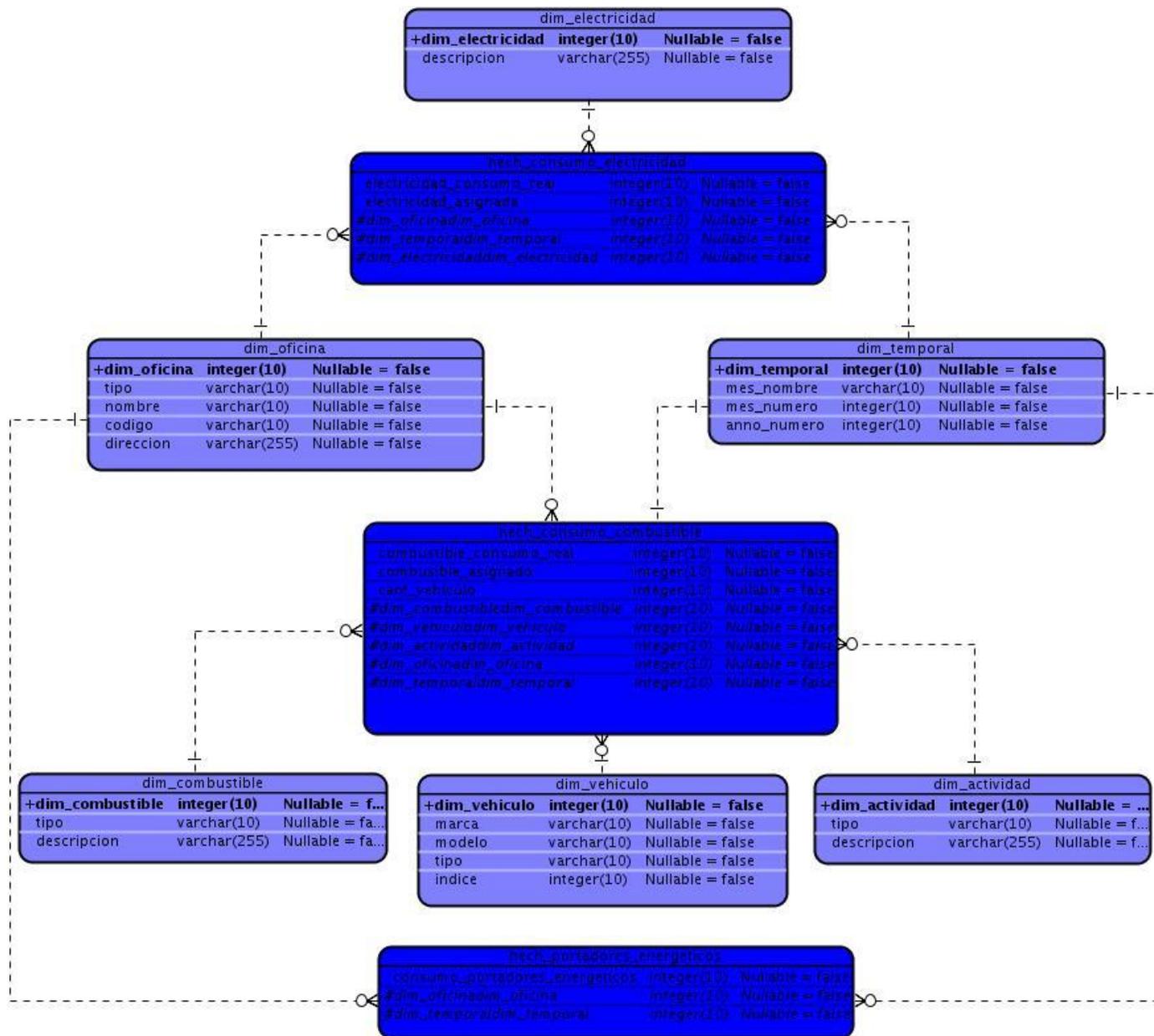
Scripts: Es un conjunto de instrucciones que permiten la automatización de tareas, creando pequeñas utilidades.

Anexos

Anexo 1: Diseño del Modelo Lógico.



Anexo 2: Diseño Modelo Físico.



Anexo 3: Tablas existentes de la BD

Tablas	Esquemas	Propietarios	Descripción
dim_combustible	dimensiones	Especialistas	Esta tabla fue generada a partir de la dimensión combustible.
dim_electricidad	dimensiones	Especialistas	Esta tabla fue generada a partir de la dimensión electricidad.
dim_temporal	dimensiones	Especialistas	Esta tabla fue generada a partir de la dimensión temporal.
dim_oficina	dimensiones	Especialistas	Esta tabla fue generada a partir de la dimensión oficina.
dim_actividad	dimensiones	Especialistas	Esta tabla fue generada a partir de la dimensión actividad.
dim_vehiculo	dimensiones	Especialistas	Esta tabla fue generada a partir de la dimensión vehículo
hech_combustible	combustible	Especialistas	Esta tabla fue generada a partir del hecho combustible
hech_electricidad	electricidad	Especialistas	Esta tabla fue generada a partir del hecho electricidad
hech_portadores_energeticos	general	Especialistas	Esta tabla fue generada a partir del hecho portadores energéticos

A continuación se abordara más detalladamente cada una de estas tablas:

Tabla Actividad:

Llave Primaria	Nombres Atributos	Tipo Datos Atributos	No Null	Único	Descripción
x	dim_actividad	serial	x	x	Llave primaria.
	Tipo	varchar(10)	x		Tipo de clasificación que puede tener una actividad, servicio, administrativo.
	descripción	varchar(20)	x		Descripción verbal de la actividad.

Tabla Combustible:

Llave Primaria	Nombres Atributos	Tipo Datos Atributos	No Null	Único	Descripción
x	dim_combustible	serial	x	x	Llave primaria.
	tipo	varchar(10)	x		Tipo de clasificación de los combustibles, grasa, lubricantes, diesel, gasolina.
	descripción	varchar(20)	x		Descripción verbal del combustible.

Tabla Electricidad:

Llave Primaria	Nombres Atributos	Tipo Datos Atributos	No Null	Único	Descripción
x	dim_electricidad	serial	x	x	Llave primaria.
	descripción	varchar(20)	x		Descripción verbal de la electricidad.

Tabla Oficina:

Llave Primaria	Nombres Atributos	Tipo Datos Atributos	No Null	Único	Descripción
x	dim_oficina	serial	x	x	Llave primaria.
	nombre	varchar(10)	x		Nombre de la oficina.
	código	varchar(10)	x		Código de la oficina.
	nivel	varchar(10)	x		Tipo de clasificación de las oficinas, nacional, provincial, municipal.
	dirección	varchar(20)	x		Dirección particular de las oficinas.

Tabla Temporal:

Llave Primaria	Nombres Atributos	Tipo Datos Atributos	No Null	Único	Descripción
----------------	-------------------	----------------------	---------	-------	-------------

x	dim_temporal	serial	x	x	Llave primaria.
	mes_nombre	varchar(10)	x		Nombre del mes.
	mes_numero	integer	x		Número del mes.
	anno_numero	integer	x		Número del año.

Tabla Vehículo:

Llave Primaria	Nombres Atributos	Tipo Datos Atributos	No Null	Único	Descripción
x	dim_vehiculo	serial	x	x	Llave primaria.
	marca	varchar(10)	x		Marca del vehículo.
	modelo	varchar(10)	x		Modelo del vehículo.
	tipo	varchar(10)	x		Tipo de vehículo.
	índice	integer	x		Índice de consumo de combustible del vehículo

Tabla Consumo de Combustible:

Llave Primaria	Llave Foránea	Nombres Atributos	Tipo Datos Atributos	No Null	Descripción
x	x	dim_oficina	integer	x	Llave foránea.
x	x	dim_actividad	integer	x	Llave foránea.
x	x	dim_combustible	integer	x	Llave foránea.
x	x	dim_temporal	integer	x	Llave foránea.
x	x	dim_vehiculo	integer	x	Llave foránea.
		consumo_real	integer	x	Consumo real de combustible.
		consumo_asignado	integer	x	Consumo asignado por oficina
		cant_vehiculos	integer	x	Cantidad de vehículo de cada

		oficina.
--	--	----------

Tabla Consumo de Electricidad:

Llave Primaria	Llave Foránea	Nombres Atributos	Tipo Datos Atributos	No Null	Descripción
x	x	dim_oficina	integer	x	Llave foránea.
x	x	dim_temporal	integer	x	Llave foránea.
x	x	dim_electricidad	integer	x	Llave foránea.
		consumo_real	integer	x	Consumo real de electricidad.
		consumo_asignado	integer	x	Consumo asignado por oficina

Tabla Portadores Energéticos:

Llave Primaria	Llave Foránea	Nombres Atributos	Tipo Datos Atributos	No Null	Descripción
x	x	dim_oficina	integer	x	Llave foránea.
x	x	dim_temporal	integer	x	Llave foránea.
		consumo_portadores_energeticos	integer	x	Consumo general asignado por oficina sin importar el tipo de portador.