Facultad Regional “Mártires de Artemisa”

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Título: *Mercado de Datos para la Dirección General Infraestructura y Servicios de la Administración Provincial de Artemisa.*

Autoras

Sonia Fernández Henríquez

Diana Rosa Prieto del Rio

Tutora:

Ing. Ana Niuska Navarro Rodríguez.

Co-Tutor:

Msc. Gilberto Mascareño Pérez

Artemisa, junio de 2012

DECLARACIÓN DE AUTORÍA

Declaramos ser las únicas autoras de este trabajo y autorizamos a la Facultad Regional “Mártires de Artemisa” de la Universidad de las Ciencias Informáticas; así como a dicho centro para que hagan el uso que estimen pertinente con este trabajo.

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

Autora: _____
Sonia Fernández Henríquez

Tutora: _____
Ing. Ana Niuska Navarro Rodríguez

Autora: _____
Diana Rosa Prieto del Rio

Co-Tutor: _____
Msc. Gilberto Mascareño Pérez

"En tiempos de cambio, quienes estén abiertos al aprendizaje se adueñarán del futuro, mientras que aquellos que creen saberlo todo estarán bien equipados para un mundo que ya no existe"

Eric Hoffer

"Donde hay una empresa de éxito, alguien tomó alguna vez una decisión valiente"

Peter Drucker

Agradecimientos

Sonia

A mis padres y a mi hermano por darme aliento siempre que me hizo falta, por su apoyo incondicional.

A Monchy por su disposición en aclarar nuestras dudas, por su tiempo, porque sin él este trabajo no se hubiera terminado.

A Ana Niuska nuestra tutora por su disposición, por guiarnos en todo momento, por brindarnos su amistad.

A Diana por siempre confiar en mí, por animarme siempre, por su perseverancia en la realización de este trabajo, por ser una amiga insuperable.

A Marla por aportar su granito de arena a que la realización de este trabajo saliera lo mejor posible, por su amistad incondicional.

Al equipo de desarrollo: Aylen y Adilen por todos sus consejos, y la ayuda brindada.

A Luis Ernesto por su paciencia, por estar siempre que me hizo falta, por sus buenas ideas.

A Rosy por sus criterios siempre acertados, por su cariño.

A todas mis amistades en especial: Yaicel, Lili, Marta, Carlos, Reinier, Pititi, por sus consejos, por ayudarme siempre que pudieron.

A todas aquellas personas que me apoyaron durante toda la carrera y han hecho posible este trabajo.

Diana Rosa

A mi mamá por sus valiosas recomendaciones y puntos de vistas.

Al profesor Monchy por su ayuda incondicional.

A Marla por ser la luz en todo momento de la realización de esta tesis y por hacerme crecer, por ser una verdadera amiga.

A Sonia por su ilimitado empeño en la realización de la tesis y por confiar en mí para hacer de nuevo una tesis, por ser una amiga sin igual.

A Aylen y Adilen por ser parte del equipo y trabajar juntas.

A Carlos por soportarme cuando las cosas no me salían y por ayudarme en todo lo que pudo.

A Luis Ernesto por su creatividad y buenas ideas.

A nuestra tutora por la confianza que nos dio sin conocernos.

A las muchachitas de mi cuarto por sus ideas en todos los cortes y por todos los momentos que hemos vivido en estos cinco años.

A todos los estudiantes del proyecto que aclararon nuestras dudas.

A los profesores de MIC a pesar de las discrepancias.

Dedicatoria

A nuestros padres,
por sembrar en nosotras el amor
y por enseñarnos a que debemos
tener la fortaleza de continuar
hacia delante no importa las circunstancias
que la vida nos presenta.

Resumen

El presente trabajo de diploma se enmarca en el tema de los Almacenes y Mercados de Datos y surge a partir de la necesidad de la Dirección General Infraestructura y Servicios de la Administración Provincial (AP) de Artemisa de contar con un sistema para el apoyo a la toma de decisiones. El objetivo que se persigue es desarrollar un Mercado de Datos que después de implementado logrará un mejor aprovechamiento de la información disponible en esta Dirección. En la investigación se detallan las metodologías, herramientas y tendencias actuales para el desarrollo de este tipo de soluciones. Además se recoge el Análisis, Diseño e Implementación del antes mencionado Mercado de Datos. Como resultado se obtiene un Mercado de Datos poblado y una capa de visualización de los datos.

ÍNDICE

ÍNDICE DE IMÁGENES

INTRODUCCIÓN

Actualmente en las empresas se necesita tomar decisiones en respuesta a las exigencias de las actividades y funciones que sus trabajadores realizan. Se hace necesario mantener los recursos de información inteligentemente organizados, integrados y resumidos. Grandes volúmenes de datos muchas veces provoca falta de concentración y homogeneización de los mismos, por lo que resulta difícil obtener una visión global del comportamiento del negocio. Hoy en día las soluciones desarrolladas en el campo de la Inteligencia de Negocios están enfocadas a lograr un mejor aprovechamiento de la información disponible en la organización, potenciando mayores niveles de seguridad y estabilidad. Uno de los resultados más significativos en este campo, ha sido la concepción de una nueva manera destinada a apoyar los procesos de toma de decisiones: los Almacenes de Datos, que representan un avance en la tecnología de las bases de datos y con ellos se pretende consolidar, integrar y centralizar los datos que las entidades generan. La unión entre el mundo de los datos y el de los negocios, por medio de la Inteligencia de Negocios con una solución basada en Almacenes de Datos permite utilizar los datos operativos de una empresa para producir información relevante para las mismas.

Las principales esferas de la sociedad en el mundo aplican la tecnología de los Almacenes de Datos con el fin de obtener de ellos la mayor información posible. Estos constituyen un instrumento útil para los directivos de las instituciones, pues le permiten formular preguntas, realizar consultas y analizar los datos según sus necesidades, sin tener que acudir al personal informático de la institución. Particularmente los Almacenes de Datos orientados al análisis de datos estadísticos se han convertido en herramientas que ofrecen beneficios para llevar a cabo los diferentes procesos dentro de las organizaciones, constituyendo uno de los principales eslabones para la toma de decisiones.

En Cuba debido al férreo bloqueo impuesto por EEUU hay un cierto atraso en el uso de las Tecnologías de la Información y las Comunicaciones (TIC), tanto en la

utilización de los diferentes software como también en el hardware, a pesar de esto el Estado dedica considerables esfuerzos para que el uso de las mismas sea masivo y abarcador, además de cuantiosos recursos en la adquisición de dispositivos informáticos y en la preparación de profesionales de este sector. Las soluciones que incluyen Inteligencia de Negocios necesitan herramientas que en su mayoría son propietarias, siendo los EEUU uno de los principales proveedores de estas tecnologías, sin embargo el país ha buscado alternativas de software libre para la implantación de estas soluciones. Desde hace varios años las empresas cubanas se encuentran sumergidas en el proceso de informatización, obteniendo mayor ventaja en la gestión y el análisis de la información.

En el país existe aún desconocimiento en torno al tema de Inteligencia de Negocios y los Almacenes de Datos, ya que en varias organizaciones no se tiene plena conciencia de las funcionalidades que estas tecnologías pueden proporcionar. A pesar de esto con el paso del tiempo las entidades se van sintiendo cada vez más atraídas por los beneficios asociados a las soluciones de Inteligencia de Negocios.

Es importante destacar que las organizaciones cubanas precisan de una mayor automatización para mejorar el desempeño de sus tareas operativas pues en algunos casos existe poca homogeneización de la información disponible a este nivel.

La Universidad de las Ciencias Informáticas (UCI) con su nuevo modelo de formación de ingenieros comprometidos con la Revolución, capaces de asimilar rápidamente los cambios tecnológicos que tienen lugar en el avance vertiginoso de la informática. Tiene como objetivo impulsar el desarrollo económico y la informatización del país. La universidad está compuesta por diferentes facultades las cuales están asociadas a centros de desarrollo de software que tienen convenios con organismos y entidades tanto nacionales como internacionales.

Entre la Facultad Regional “Mártires de Artemisa” de la UCI y la Administración Provincial de Artemisa surgió un convenio en el marco de la distribución política-

administrativa del país realizada a inicios del año 2011, con el propósito de obtener un incremento en la eficiencia de los procesos que desarrolla cada dirección que la conforma.

La AP de Artemisa cuenta con 32 direcciones, específicamente este trabajo está relacionado con la Dirección General Infraestructura y Servicios, integrada por las direcciones provinciales de Comercio, Recursos hidráulicos, Transporte y Vivienda, y tiene como funciones principales:

- Dirigir el proceso de determinación de los planes de conservación y obras nuevas así como el proceso inversionista aprobado para el patrimonio de la AP.
- Dirigir la actividad de transporte de cargas y pasajeros, sus servicios auxiliares y conexos y el mantenimiento de la infraestructura técnica.
- Dirigir el cumplimiento de la política del Estado con relación al patrimonio de los recursos hídricos e hidráulicos así como su instrumentación y cumplimiento.
- Dirigir el comercio minorista de bienes de consumo, comercio mayorista y logística de almacenes en los sectores estatal, cooperativo, mixto y privado que operan en el territorio.

Actualmente en esta Dirección se dificulta el análisis de los indicadores a medir en los procesos que se llevan a cabo, debido a la gran cantidad de información descentralizada. Los datos presentan inconsistencias, debido al uso de diferente codificación en la misma información. La información necesaria no está disponible para la realización de informes inmediatos. La elaboración de informes resulta costosa en esfuerzo y tiempo debido al gran cúmulo de información.

Las deficiencias descritas anteriormente traen como consecuencia que se vean afectadas la centralización, consistencia y disponibilidad de la información haciendo

el análisis de ésta complicado y lento, lo que dificulta el proceso de la toma de decisiones.

De las situaciones anteriormente mencionadas surge el siguiente **problema a resolver**: ¿Cómo contribuir a la centralización, consistencia y disponibilidad de la información para facilitar el proceso de la toma de decisiones en la Dirección General Infraestructura y Servicios de la Administración Provincial de Artemisa?

Definiendo como **objeto de estudio**: Los Almacenes de Datos, y como **campo de acción**: Mercados de Datos en los sectores del comercio, el transporte, la vivienda y los recursos hidráulicos.

Para dar solución a la problemática que dio surgimiento al presente trabajo se ha propuesto como **objetivo general**: Desarrollar un Mercado de Datos para contribuir a la centralización, consistencia y disponibilidad de la información para facilitar el proceso de la toma de decisiones en la Dirección General Infraestructura y Servicios de la Administración Provincial de Artemisa.

A partir del objetivo trazado se asumen las siguientes **preguntas científicas**:
¿Cuáles son los criterios teóricos relacionados con los Almacenes de Datos?
¿Cómo se realiza el proceso de análisis de la información en la Dirección General Infraestructura y Servicios?
¿Cómo desarrollar el Mercado de Datos Infraestructura y Servicios?

Para darle respuestas a las interrogantes planteadas se determinaron las siguientes **tareas de la investigación**:

1. Fundamentación del marco teórico de los Almacenes de Datos.
2. Caracterización del proceso de análisis de la información en la Dirección General Infraestructura y Servicios.
3. Desarrollo del Mercado de Datos Infraestructura y Servicios.
4. Validación de las funcionalidades del Mercado de Datos Infraestructura y Servicios.

Para el desarrollo de la solución se emplearon **métodos científicos de la investigación Teóricos y Empíricos**.

Métodos Teóricos

Análisis Histórico-Lógico: Este método permitió investigar el funcionamiento y desarrollo de los elementos útiles en la investigación, así como su comportamiento en el transcurso de la historia.

Análisis-Síntesis: Este método permitió estudiar el proceso de almacenamiento de la información para la toma de decisiones y buscar información referente a los tipos de gestores de base de datos existentes, así como las herramientas y métodos que se utilizan para el desarrollo del Almacén de Datos.

Modelación: Es considerado de gran importancia para el diseño del Mercado de Datos, por lo que se utilizó para el modelado del negocio, y el modelado dimensional, el cual demostró con claridad la forma en que se va a presentar la solución.

Métodos Empíricos

Para la aplicación de los métodos empíricos en esta investigación se contó con todos los trabajadores de la Dirección General Infraestructura y Servicios y los trabajadores de cada una de las Direcciones Provinciales que la integran, constituyendo la mejor fuente de información.

Entrevista: Este método permitió conocer la forma en que se maneja la información en la Dirección General Infraestructura y Servicios mediante una entrevista a los trabajadores de la misma. (Ver Anexo 1)

Observación: Este método permitió recoger información del tratamiento de los datos en la Dirección General Infraestructura y Servicios, obteniendo así conocimientos del tema a partir de situaciones dadas.

El **aporte práctico** de la investigación es un Mercado de Datos para la Dirección General Infraestructura y Servicios.

Para cumplir con todos los elementos planteados, el presente trabajo está estructurado en cuatro capítulos:

Capítulo 1: Fundamentación teórica de los Almacenes Datos. Se realiza un estudio del estado del arte de los Almacenes y Mercados de Datos existentes en el mundo y en el país. También se aborda sobre las principales definiciones, metodologías y herramientas que se utilizarán para el desarrollo de la solución.

Capítulo 2: Análisis y Diseño del Mercado de Datos Infraestructura y Servicios. Se definen los requerimientos del sistema, se diseña el diagrama de Casos de Uso del Sistema (CUS), el modelo de datos, los subsistemas de integración y visualización. En este capítulo se describe la arquitectura del Mercado de Datos.

Capítulo 3: Implementación del Mercado de Datos Infraestructura y Servicios. Se implementan los procesos de extracción, transformación y finalmente la carga al Mercado de Datos, y también la implementación del subsistema de visualización con el objetivo de darle solución a los requisitos del sistema.

Capítulo 4: Validación de las funcionalidades del Mercado de Datos Infraestructura y Servicios. Se prueba el Mercado de Datos a través de los casos de prueba y las pruebas de integración con el objetivo de encontrar y documentar los defectos que pueden afectar la calidad del sistema.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA DE LOS ALMACENES DATOS.

Introducción

En este capítulo se hace un estudio de los Almacenes y los Mercados de Datos, sus características, los elementos que lo componen y metodologías de desarrollo. Se realiza además un estudio del estado del arte tanto a nivel mundial como nacional sobre la utilización de esta tecnología. Se aborda también sobre la integración de los datos y la Inteligencia de Negocios, así como las principales herramientas para el desarrollo de los Almacenes de Datos.

1.1 Almacén de Datos

En la actualidad con el creciente aumento de la información, las bases de datos se han convertido en una de las herramientas más útiles en cuanto a la gestión de la información a nivel científico, social, económico, político y cultural.

El cuantioso número de información que acumulan las grandes empresas y negocios trajo consigo la necesidad de darle un fin útil a toda esta acumulación de información, donde estaba almacenada la mayor parte de las operaciones de las mismas.

De esta manera la Ciencia de la Información, ha desarrollado una producción científica importante a nivel mundial, la cual ha utilizado las bases de datos, como repositorio de almacenamiento y difusión de información. (Soto, 2001)

Este aumento de los volúmenes de la información y del nivel competitivo de las empresas ha dado lugar al desarrollo de nuevas estrategias de gestión de la información, lo cual conduce al concepto de Almacén de Datos.

Son muchos los conceptos entorno a la tecnología de los Almacenes de Datos, para comprenderla se hace necesario conocer algunas definiciones.

Según Inmon los Almacenes de Datos son, “un conjunto de datos orientados a un tema, integrados, de tiempo variante y no volátiles usados en la estrategia de toma de decisiones administrativas.” Por su parte Ralph Kimball, plantea que“…los Data

Warehouses son una copia de los datos de la transacción estructurados específicamente para la pregunta y el análisis.” (Rivadera, 2010)

Es así que un Almacén de Datos es una colección de datos históricos, para dar soporte a la toma de decisiones.

1.2 Mercados de Datos

Los Mercados de Datos, son un subconjunto de datos de un Almacén de Datos donde se almacenan la mayoría de las actividades de análisis que en el entorno de Inteligencia de Negocios se llevará a cabo. (Sierra, 2009)

Los Mercados de Datos se ajustan a la definición de un Almacén de Datos, con la diferencia de que los primeros sirven a un área específica del negocio. Los procesos y fuentes de datos necesarios para construir un Mercado de Datos son iguales que en el caso del Almacén, pero estos son más pequeños y la información almacenada se obtiene de un menor número de fuentes.

Los Almacenes y los Mercados de Datos resultan ser una base de datos centralizada que contienen datos de una o diversas fuentes, orientados a permitir el análisis detallado de los datos.

A partir de lo planteado anteriormente los Mercados de Datos, como estructura, son un Almacén de Datos pero reducido a un departamento específico.

Características de los Almacenes de Datos

Los Almacenes de Datos tienen cuatro características fundamentales:

- Temático: El Almacén de Datos debe contener información específica sobre los principales temas o prioridades de la organización.
- Integrado: Los datos almacenados deben integrarse en una estructura consistente, esto se refiere a que el valor de un mismo atributo debe estar representado de igual forma.
- Histórico: Los datos se almacenan por períodos de tiempo lo que permite hacer comparaciones sobre el estado de los mismos.

- No volátil: La información que es almacenada solo es leída, no puede ser modificada. Un Almacén de Datos no soporta las operaciones de insertar, modificar y eliminar como lo hace una base de datos operacional.

Componentes de los Almacenes de Datos

Los elementos por los que están compuestos los Almacenes de Datos definen el ambiente que estos poseen.

Independientemente de que cada proceso de desarrollo de un Almacén de Datos es único, debido a que se siguen las especificidades de cada empresa u organización, la mayoría de las veces cumplen con la ejecución de los componentes de un Almacén de Datos.

La siguiente imagen muestra los componentes del Almacén de Datos y la relación entre ellos: (Sierra, 2009)

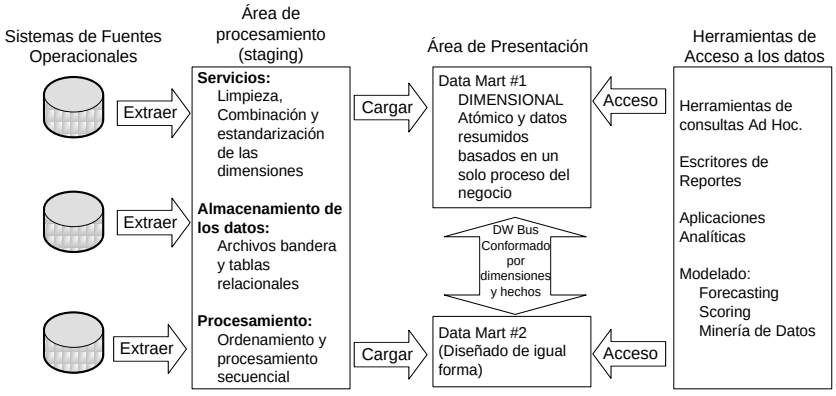


Ilustración 1 Relación entre los componentes de un Almacén de Datos

Sistema de fuentes operacionales

Son los sistemas con los que cuentan las empresas para la gestión de su flujo de información diario, se almacenan en distintos formatos y tipos de ficheros, en los cuales se puedan hacer cualquier tipo de consultas.

Sobre estas fuentes se tiene poco o ningún control sobre el volumen y formato de los datos, teniendo como prioridades principales el procesamiento, el rendimiento y la disponibilidad. Generalmente realizan salvas de la información que gestionan y

sólo trabajan con los datos generados en un período corto de tiempo para hacer las recuperaciones de forma más óptima. También existe la posibilidad de que sean fuentes creadas manualmente debido a que no posean un sistema que las procese.

Área de procesamiento

Es el área donde se realizan los procesos de Extracción, Transformación y Carga (ETL, por sus siglas en inglés) y en donde se almacenan los datos temporalmente.

En este componente se realiza la extracción de los datos de las diversas fuentes operacionales que se vayan a integrar, donde su principal tarea es almacenar toda esa información en bases de datos relacionales, generalmente, para realizar el análisis y procesamiento de los datos. Cuando los datos son almacenados en bases de datos temporales se lleva a cabo el proceso de limpieza, el cual consiste en eliminar todos los errores e inconsistencias, con el objetivo de lograr la estandarización de los datos para luego cargarlos en el Almacén. (Ponniah, 2001)

Área de presentación

En este componente se ubica la información que es relevante para los usuarios a la hora de la toma de decisiones. Los datos se encuentran organizados, almacenados y disponibles para cualquier tipo de análisis que se solicite por parte de los usuarios.

Esta área se encuentra la mayoría de las veces representada por un conjunto de Mercados de Datos integrados, en donde cada uno representa un proceso del negocio específico.

Herramientas de acceso a datos

En este componente se usa la palabra herramientas para referirse a la variedad de capacidades que pueden ser provistos a los usuarios del negocio para el soporte a la toma de decisiones. Su actividad principal es consultar el área de presentación del Almacén de Datos. El mismo puede abarcar desde una simple o personalizada herramienta de consulta, hasta una compleja y sofisticada aplicación de modelado o minería de datos. (Ponniah, 2001)

1.3 Modelo dimensional

Los Almacenes de Datos se diseñan mediante el modelo dimensional, la información está organizada de forma tal que garantiza la velocidad y eficiencia en la recuperación de la misma. Este modelo se compone por hechos, dimensiones y medidas. Para modelar la estructura de un Almacén de Datos se utilizan los siguientes esquemas dimensionales:

Esquema en estrella

En el esquema en estrella la tabla de hechos es la única tabla que tiene múltiples joins que la conectan con otras tablas. El resto de tablas del esquema (tablas de dimensión) únicamente hacen join con esta tabla de hechos. Toda la información referente a una dimensión se almacena en la misma tabla. (Sanz, 2010)

Esquema copo de nieve

En cada dimensión se almacenan jerarquías de atributos o bien simplemente se separan atributos en otra entidad por razones de desempeño y mejor utilización del espacio. El uso más común de este esquema es cuando las tablas de dimensiones son muy grandes o complejas y es muy difícil representar los datos en un esquema estrella. (Mendez, 2009)

Constelación de hechos

Para cada esquema estrella o esquema copo de nieve de un Almacén de Datos es posible construir un esquema de constelación de hechos. Este esquema es más complejo que las otras arquitecturas debido a que contiene múltiples tablas de hechos. Con esta solución las tablas de dimensiones pueden estar compartidas entre más de una tabla de hechos. (ETL-Tools.Info, 2006)

Tablas de dimensiones

Las tablas de dimensiones complementan las tablas de hechos. La mayor parte de las dimensiones contienen un gran número de atributos de texto que sirven de base para restringir y agrupar las consultas sobre el Almacén de Datos. (Espinosa, 2007)

Tablas de hechos

Las tablas de hechos son las tablas principales en el modelo dimensional y contiene las mediciones sobre los atributos de la organización, valores del negocio. Los hechos más útiles son los que contienen información numérica y aditiva. (Espinosa, 2007)

Cada una de las mediciones es tomada como la intersección de todas las dimensiones. Cada fila de esta tabla corresponde a un hecho determinado y a su vez cada conjunto de hechos dentro de la tabla de hechos referencian a la misma granularidad.

Medidas

Una medida es un atributo (campo) de una tabla que se desea analizar, sumarizando o agrupando sus datos, usando los criterios de corte conocidos como dimensiones. (Rivadera, 2010)

Las medidas más útiles para incluir en una tabla de hechos son aquellas medidas numéricas que pueden calcularse con la suma de varias cantidades de la tabla. Es decir, las medidas candidatas son los datos numéricos, pero no cada atributo numérico es una medida candidata. En consecuencia, por lo general los hechos a almacenar en una tabla de hechos van a ser casi siempre valores numéricos, enteros o reales.

Granularidad

La granularidad significa especificar el nivel de detalle. La elección de la granularidad depende de los requerimientos del negocio y lo que es posible a partir de los datos actuales. (Rivadera, 2010)

La granularidad de la tabla de hechos representa el nivel más atómico por el cual se definen los datos.

Atributos e indicadores

Los atributos son criterios utilizados para analizar los indicadores. Se basan, en los datos de referencia de las tablas de dimensiones. Son campos o criterios de análisis, pertenecientes a las tablas de dimensiones. (González, 2011)

La calidad de todo Almacén de Datos se mide por la definición de los atributos de las dimensiones. Su poder es directamente proporcional a la calidad y profundidad de estos atributos. (Kimball, 2008)

Los indicadores son variables que pueden tomar un valor de una determinada unidad de medida y de un determinado tipo de datos. (González, 2011)

1.4 Modos de almacenamiento de datos

El Proceso Analítico en Línea (OLAP, por sus siglas en inglés) define a una tecnología que se basa en el análisis multidimensional de los datos y que le permite al usuario tener una visión más rápida e interactiva de los mismos. (Lanzillotta, 2012)

Utilizando herramientas OLAP, los usuarios pueden acceder al Almacén de Datos, brindando a los responsables de la toma de decisiones en las organizaciones el potencial de mejorar su comprensión del negocio.

Existen tres modelos para el Proceso Analítico en Línea de la información: ROLAP, MOLAP y HOLAP, dependiendo de las técnicas que se utilicen a la hora de obtener los datos y la forma en la que están estructurados. El proceso de análisis se realiza de igual forma, variando solo entre uno y otro caso 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.

ROLAP

En el Procesamiento Analítico Relacional en Línea (ROLAP, por sus siglas en inglés) se accede a los datos almacenados en un Almacén de Datos para

proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales. (Nader, 2003)

Además los datos son almacenados en filas y columnas de forma relacional, se crean vistas multidimensionales extrayendo los datos de bases de datos SQL relacionales. Este modelo presenta los datos a los usuarios en forma de dimensiones de negocio.

La arquitectura ROLAP es capaz de usar datos precalculados si estos están disponibles, o de generar dinámicamente los resultados desde los datos elementales si es preciso.

MOLAP

El Procesamiento Analítico Multidimensional en Línea (MOLAP, por sus siglas en inglés) usa bases de datos multidimensionales para proporcionar el análisis. Aquí las estructuras de los datos están fijas para que la lógica al procesar el análisis multidimensional pueda ser basada en métodos bien definidos para establecer las coordenadas del almacenamiento de los datos. Las herramientas MOLAP acceden a datos que no están almacenados en registros de tablas, sino que almacenan los datos en arreglos de varias dimensiones, llamados cubos. Estos cubos utilizan índices para optimizar el acceso a los datos, provocando que el acceso a la información almacenada se realice de forma más rápida y efectiva utilizándose en Almacenes de Datos donde el tiempo en la velocidad de respuesta es crítico. (Business Intelligence+Informática estratégica, 2002)

HOLAP

El modo de almacenamiento HOLAP (HOLAP, por sus siglas en inglés), combina las arquitecturas ROLAP y MOLAP para brindar una solución con las mejores características de ambas: desempeño superior y gran escalabilidad. Un tipo de HOLAP mantiene los registros de detalle en la base de datos relacional, mientras

que mantiene las agregaciones en un almacén MOLAP separado. (Business Intelligence+Informática estratégica, 2002)

En relación con todo lo mencionado anteriormente se escoge como modo de almacenamiento de datos al ROLAP.

1.5 Integración de datos

El aspecto más importante del ambiente del almacenamiento de datos es que la información esté siempre integrada.

La desintegración de los datos muchas veces afecta la calidad de la información. El proceso de integración se basa en la necesidad de unir los datos provenientes de diferentes fuentes con el objetivo de centralizar toda la información para un mejor análisis de la misma. Además se busca la estandarización de los datos y la consistencia de la información.

Integración de los datos. Características

El proceso de Extraer, Transformar y Cargar se refiere a los datos en una empresa. ETL es el proceso que organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un Almacén de Datos, limpiarlos y cargarlos en otra base de datos, Mercado de Datos o Almacén de Datos. (ETL-Tools.Info, 2006)

Proceso de Extracción Transformación y Carga

Los procesos de ETL constan de múltiples pasos, cuyo objetivo es transferir datos desde las aplicaciones de producción a los sistemas de Inteligencia de Negocios:

- Extracción de los datos desde las distintas fuentes tanto internas como externas. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación.
- Transformación de estos datos para obtenerlos lo más precisos, completos, consistentes, interpretables y accesibles. Después del proceso de limpieza se lleva a cabo la integración de los datos con el propósito de eliminar

problemas de redundancia e identificar las fuentes de datos más fiables. Una vez realizado el proceso de extracción y limpieza se procede a transformar los datos para de esta forma estandarizar los códigos, corregir los datos, eliminar registros duplicados, usar conversiones y combinaciones para generar nuevos campos.

- Carga de los datos resultantes en las diversas aplicaciones de Inteligencia de Negocios. En este paso se organizan y actualizan los datos y los metadatos en la base de datos.

Si no se realiza un correcto proceso de ETL se pudieran obtener datos incorrectos lo que afectaría el proceso de toma de decisiones, es por eso que este proceso constituye aproximadamente un 70% del trabajo de la construcción de un Almacén de Datos. (Business Intelligence+Informática estratégica , 2007).

1.6 Inteligencia de Negocios

Se le llama Inteligencia de Negocios al conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existente en una organización. (Cañete, 2006)

De acuerdo al Data Warehousing Institute, la definición de Inteligencia de Negocios es la siguiente: "…Son los procesos, tecnologías, y herramientas que se necesitan para convertir los datos en información, la información en conocimiento, y el conocimiento en planes que impulsan acciones rentables para el negocio. La Inteligencia de Negocios abarca el almacenamiento de datos, herramientas analíticas, y contenido y gestión del conocimiento…" (Armstrong, 2006.)

La Inteligencia de Negocios apoya a los encargados de tomar decisiones con la información correcta, en el momento y lugar correcto, lo que les permite tomar mejores decisiones de negocios. La información adecuada en el lugar y momento adecuado incrementa la efectividad de cualquier empresa.

Las soluciones de Inteligencia de Negocios reportan numerosas ventajas sobre todo a las empresas a elevar su nivel competitivo en el mercado, facilitando el análisis de información útil a las organizaciones para la toma de decisiones. Asiste a los ejecutivos para planear y pronosticar el trabajo, presentando una descripción común de los procesos del negocio de una compañía.

Entre las principales ventajas que ofrece una solución de Inteligencia de Negocios se tienen: (NEXTEL Engineering, 2010)

- Control de costes, al tener un solo sistema que permite manejar fácilmente los distintos programas que se encuentran en los diferentes departamentos de su compañía. Mejora de la colaboración y la calidad de las decisiones, facilitando el acceso a la información en todos los niveles de la organización.
- Orienta las soluciones tecnológicas hacia el usuario, porque reduce los tiempos de aprendizaje mediante el uso de herramientas de uso cotidiano.
- Proporciona una profunda visión del negocio a través de un sistema integrado de usos: Cuadros de Mando Integrales, Dashboards Digital, Consultas y Reportes, Minería de Datos, y Almacenamiento Analítico.
- Asiste a los ejecutivos para planear y pronosticar el trabajo, presentando una descripción común de los procesos del negocio de una compañía.

1.7 Metodologías para el desarrollo

Existen muchas metodologías para el diseño y la construcción de un Almacén de Datos destacándose fundamentalmente dos, la metodología de Kimball y la de Inmon.

Bill Inmon plantea que la construcción de un Almacén de Datos comienza por el Almacén en sí. En su filosofía los Mercados de Datos son dependientes del Almacén de Datos y por tanto se construyen después de él. Se puede decir además que Inmon es defensor de utilizar el modelo relacional para el ambiente en el que se desarrollará el Almacén, afirmando que la creación de una base de datos relacional con una ligera normalización, son la base de los Mercados de Datos. O lo

que es lo mismo, a partir de los esquemas relacionales, a los que se les irá añadiendo complejidad, se obtendrán finalmente los Mercados de Datos.

La metodología de Inmon no viene acompañada del ciclo de vida normal de las aplicaciones. En esta metodología los requisitos irán acompañando al proyecto según vaya comprobándose su necesidad. Esta visión puede traer consigo mucho riesgo a la compañía, ya que invierte grandes esfuerzos en el desarrollo del Almacén y no es hasta la aparición de los Mercados cuando se empieza a explotar la inversión y obtener beneficios. Esta metodología implica un gran consumo de tiempo, es por esto que algunas empresas se inclinan por metodologías con las que obtengan resultados tangibles en un espacio menor de tiempo.

Por su parte Ralph Kimball plantea que se debe partir de la construcción de los Mercados de Datos primero y que el conjunto de estos es lo que conformará el Almacén de Datos, lo que conduce a una solución en una cantidad pequeña de tiempo.

Las principales diferencias de estas dos metodologías están basadas en la presentación de la información y el tratamiento de la información atómica. La mayoría de las empresas se inclinan por la metodología de Kimball puesto que generalmente se parte por un departamento dentro de la empresa, lo que se convierte en un Mercado de Datos, sumándose más tarde la construcción de más Mercados de Datos para luego formar un Almacén de Datos.

Tomando como base estas metodologías se han desarrollado otras que no siguen obligatoriamente una específica sino que realizan una selección de lo mejor de cada una y definen su propia metodología. Una de ellas es la Metodología para el Diseño Conceptual de Almacenes de Datos, presentada en la Tesis de Doctorado de Leopoldo Zenaido Zepeda Sánchez que aporta como aspecto novedoso con respecto a las anteriores la incorporación de los Casos de Uso (CU) 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. (Sánchez, 2008)

Justificación de la metodología a utilizar

Para el desarrollo del Mercado de Datos se tomó la Metodología para el desarrollo de soluciones de Almacenes de Datos e Inteligencia de Negocios adoptada por la Universidad, la cual toma como base la metodología de Kimball por los siguientes aspectos:

- Crea los conceptos de Hechos y Dimensiones.
- Propone ir construyendo el Almacén de Datos a través de la construcción de los Mercados de Datos departamentales.
- Existe abundante documentación sobre la misma.

La metodología de desarrollo seleccionada está organizada por fases y grupos de trabajos, algunas de estas fases podrán ser implementadas de forma paralela según el componente que se está desarrollando para integrarse al final de la solución, lo que permite un desarrollo más ágil.

El ciclo de vida de esta metodología consta de las siguientes fases: Estudio Preliminar, Requerimientos, Arquitectura y Diseño, Implementación, Prueba, Despliegue y Soporte.

En la ejecución de cada fase participan un grupo de personas encargadas del desarrollo del proyecto. Los recursos humanos están organizados por grupos de trabajo por lo que los roles del proyecto están distribuidos por cuatro grupos de trabajo, donde cada uno realiza actividades específicas y bien delimitadas según sus responsabilidades en el desarrollo del proyecto. Estos grupos son: Grupo de Análisis, Grupo de Almacenes, Grupo de ETL y Grupo de BI.

1.8 Herramientas existentes

Actualmente existen múltiples herramientas que se han desarrollado con el objetivo de ayudar en la construcción, implementación y en el mantenimiento de un

Almacén de Datos. En el marco de las herramientas de código libre, se destacan las herramientas de la suite de Pentaho dedicadas a la integración de los datos y al análisis de estos, así como Talend Open Studio y dentro de las herramientas con licencia, Informática PowerCenter y BITool.

Talend, ofrece soluciones de código abierto que permiten la integración de datos, calidad de datos, y la integración de aplicaciones. Con su probado rendimiento, facilidad de uso, capacidad de ampliación y robustez, las soluciones de Talend son unas de las soluciones de integración más ampliamente utilizadas y desplegadas en el mundo. Pentaho Data Integration a su vez proporciona la Extracción de gran alcance, Transformación y Carga utilizando capacidades innovadoras, basada en un enfoque de metadatos. Con una interfaz intuitiva y gráfica de arrastrar y soltar, entorno de diseño y una arquitectura probada, escalable y basada en estándares, siendo cada vez más la elección de las organizaciones más tradicionales de ETL, de propiedad o las herramientas de integración de datos.

Informática PowerCenter establece un estándar para alta escalabilidad, alto rendimiento de software de integración de datos. Permite a la organización implementar un enfoque único para el acceso, la transformación y la entrega de datos sin tener que recurrir a la codificación manual. Por su parte BITool es una potente herramienta que permite crear complejos procesos de cargas en tan sólo unos minutos, reduciendo notablemente el tiempo y costo en programación, mantenimiento e implementación de proyectos de Inteligencia de Negocios.

En el mundo de los gestores de base de datos, se puede decir que existen tres grandes agrupaciones. En la primera los Sistemas Gestores de Base de Datos (SGBD) libres, dentro de ellos se encuentra como uno de los más relevantes PostgreSQL. La segunda agrupación son el conjunto de los gestores no libres donde resaltan como principales ORACLE, Microsoft SQLServer y MySQL. Y por último la tercera agrupación y más pequeña los gestores considerados productos

no libres y gratuitos destacándose Microsoft SQLServer Compact Edition Basic y Sybase ASE Express Edition para Linux.

Oracle es una de las compañías más reconocidas en el ámbito de los SGBD y está considerado como uno de los más completos que existen. Como principales ventajas de esta herramienta están la estabilidad y la seguridad, además de que es multiplataforma. Actualmente sufre la competencia de otras herramientas como Microsoft SQL Server de Microsoft, MySql y PostgreSQL.

A diferencia de Oracle, MySQL es un Sistema Gestor de Bases de Datos relacional. Está herramienta fue un software de código abierto, licenciado bajo la GPL. A principios de 2009 fue comprada por la compañía de software propietario SUN MicroSystem. El lenguaje de programación que utiliza es SQL (Structured Query Language). Unas de las principales desventajas que presenta es que un gran por ciento de sus utilidades no están documentadas y no es una herramienta intuitiva.

Justificación de las herramientas a utilizar

En la presente investigación se decidió utilizar en su mayoría herramientas de código libre, siguiendo la política de software libre del país, de la universidad y del proyecto para el desarrollo de Almacenes de Datos.

Herramienta de modelado

- **Visual Paradigm 6.4:** Es una herramienta UML profesional que soporta el ciclo de vida completo del desarrollo de software. Permite realizar ingeniería tanto directa como inversa. Esta herramienta soporta múltiples usuarios trabajando sobre el mismo proyecto. Esta herramienta ayuda a la construcción de aplicaciones de calidad, mejores y a un menor costo. (Company Headquarters, 2009)

  También brinda características como generación de bases de datos. Ingeniería inversa de bases de datos desde Sistemas Gestores de Bases de Datos existentes a diagramas de Entidad-Relación. Es una herramienta fácil

de instalar y actualizar y es posible realizar ingeniería inversa-código a modelo, código a diagrama.

Gestor de Base de Datos

- **PostgreSQL 9.1:** es un Sistema de Gestión de Bases de Datos Objeto-Relacional (ORDBMS). Esta herramienta presenta las siguientes características: soporta distintos tipos de datos, además del soporte para los tipos base, también soporta datos de tipo fecha, elementos gráficos, datos sobre redes, cadenas de bits. Permite la creación de tipos propios, incorpora una estructura de datos de arreglos, incorpora funciones de diversa índole: manejo de fechas, geométricas, orientadas a operaciones con redes, entre otras, permite la declaración de funciones propias, así como la definición de disparadores, incluye herencia entre tablas (aunque no entre objetos, ya que no existen), mantiene la integridad de los datos, controla el acceso concurrente de los usuarios, facilita el manejo de grandes volúmenes de información.

- **PgAdmin III 1.12:** es una aplicación gráfica para gestionar el gestor de bases de datos PostgreSQL, siendo la más completa y popular con licencia Open Source. Soporta todas las características de PostgreSQL y facilita enormemente su administración. Permite escribir consultas SQL simples, a través de un editor SQL con resaltado de sintaxis. Esta herramienta incluye también un editor de código de la parte del servidor y un agente para lanzar scripts programados.

Herramientas para la integración de datos

Las herramientas para la integración de datos son muy útiles para el proceso de ETL garantizando ganancias en tiempo y total fiabilidad de los datos.

- **Pentaho Data Integration 3.2.0:** Es una herramienta multiplataforma, tiene un entorno gráfico de desarrollo. Usa las tecnologías Java, XML, Java Script, es muy fácil de instalar y configurar. Esta herramienta está basada en dos

tipos de objetos: transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones).

- **DataCleaner 1.5.4:** Permite la evaluación del nivel de calidad de los datos contenidos en el sistema de información. Es una aplicación muy fácil de usar, genera sofisticados informes y gráficos que permiten a los usuarios determinar de un vistazo el nivel de calidad de los datos, identificar y analizar la estructura del origen de datos y combinar resultados y gráficos, creando vistas fáciles de interpretar para evaluar la calidad de los datos.

Herramientas para la Inteligencia de Negocios

Estas herramientas asisten específicamente el análisis y la presentación de los datos.

- **Schema Workbench 3.2.1:** Es una interfaz de diseño que permite crear y probar esquemas de cubos Mondrian OLAP visualmente. La Plataforma de BI de Pentaho incrusta el motor de consulta Mondrian, como parte de su arquitectura. Además permite la ejecución de consultas MDX. Está distribuida bajo la licencia EPL.
- **BI-Server 3.6:** Es una solución de Inteligencia de Negocios completa y fácil de usar. Incluye interfaces de uso autónomo con base en funciones para todo tipo de usuarios dentro de una estructura de control bien definida y un punto centralizado de administración. Esto ayuda a las empresas a simplificar y agilizar la implementación de la Inteligencia de Negocios.
- **Pentaho Metadata Editor (PME) 3.7.0**: Es una herramienta que permite crear dominios de metadatos y modelos. El objetivo es mapear la estructura física de la base de datos a un modelo lógico de negocio.
- **Pentaho Report Designer 3.7.0:** Es una herramienta de reporte que permite crear informes, ya sea para ejecutarlos directamente o para publicarlos en la plataforma de Inteligencia de Negocios y que desde allí puedan ser

utilizados por los usuarios. Incluye asistentes para facilitar la configuración de propiedades, cuenta también con un editor de consultas para facilitar la confección de los datos que serán utilizados en un informe.

- **Apache Tomcat 5.5:** Es un servidor web y de aplicaciones que gestiona solicitudes y respuestas Http (incluye el servidor Apache) gracias a sus conectores Http. Es un servidor de aplicaciones o contenedor de Servlets/JSP. Es de código abierto, implementado con tecnología. Funciona en cualquier sistema operativo en donde se encuentre la máquina virtual de Java permitiéndole a los usuarios realizar reportes.

1.9 Estado actual de los Almacenes y Mercados de Datos

Los Almacenes de Datos constituyen una nueva tecnología de la información, se consideran como un gran avance dentro de las bases de datos, convirtiéndose en una herramienta idónea para la consulta y el análisis de los datos en una empresa.

El concepto de Almacén de Datos surge a mediados de la década del 80. A partir de la década de los 90 se eleva el nivel competitivo de las empresas, obligando de cierta forma a los directivos a proponer ideas nuevas dentro del campo gerencial. A partir de entonces muchas empresas a nivel mundial han incorporado la tecnología de los Almacenes de Datos como una herramienta que les proporcione eficiencia y precisión en los análisis de la información.

El mercado minorista es uno en el que más incidencias han tenido los Almacenes de Datos. En América Latina existen empresas en el negocio de las comunicaciones que han implementado Almacenes entre ellas se destacan: TV Azteca, Wal-Mart, Visa, Telefónica de Argentina, Ipostel, GNP, Baxter. Las compañías American Stores, Canadian Tyre, Owens Corning Glass, han obtenido ya resultados tangibles en la implementación de esta nueva herramienta.

En Europa empresas como Carrefour (España) WH Smith Books (Gran Bretaña), BonPreu (España), SAR Group (España), Great Universal Gran Bretaña), Corte

Inglés (Francia), Cortefiel (Francia), Eroski (Francia), Supermercados Casino (Francia), Otto Versand (Alemania), Helene Curtis (España) han sido pioneras en el uso de esta nueva tecnología.

Las grandes transnacionales como: Coca Cola, Walt Disney, Nike, Maybelline, Adidas, 3M, Bosh Siemens se han incorporado a la utilización de los Almacenes de Datos para realizar sus estudios de mercado. Muchos de los bancos mundiales implementan esta tecnología, como es el caso del Banco de Argentina, Banfoandes, Banorte, Banco de Venezuela, Banco Nacional de España, Caja Extremadura.

En la medicina también se utiliza esta tecnología para realizar estudios de patrones de comportamiento en pacientes con diferentes patologías y para dar seguimiento a los tratamientos y a los pacientes en sí. El diario El Mundo cuenta con un Almacén de Datos cuyo objetivo es obtener información completa sobre la contratación de publicidad en sus medios.

El Instituto Nacional de Estadística e Informática (INEI) de Perú en el marco del censo realizado en el 2007 y con el objetivo de organizar y administrar la información de los censos nacionales de población y vivienda, en forma integrada y estandarizada ideó desarrollar un Almacén de Datos de los Censos Nacionales de Población y Vivienda, información económica e información social, como un medio que facilite la toma de decisiones y su monitoreo, la promoción del desarrollo sostenible social y económico, las políticas gubernamentales de focalización de la lucha contra la pobreza y el seguimiento del cumplimiento de objetivos y metas de los proyectos de desarrollo.

Por otro lado en México el Instituto Nacional de Estadística, Geografía e Informática (INEGI) tiene un Almacén integrado de datos definitivos con información estadística obtenida de los programas de censos nacionales, encuestas y registros administrativos para la elaboración de productos, la toma de decisiones y la planeación, facilitando que el personal del Instituto pueda atender con mayor

oportunidad los requerimientos de información de los usuarios del INEGI. (Salazar, 2008)

En el Centro de Investigación sobre Recursos Hídricos de la Universidad de Texas se implementó el modelo ArcHydro, como un estándar para la elaboración de bases de datos asociadas a los recursos hídricos como apoyo a la gestión. Dentro de la información espacial que se consideró están los cuerpos de agua y la ubicación de presas, estaciones hidrométricas, climatológicas y de calidad del agua, y la ubicación de aprovechamientos superficiales. Las series de tiempo incluyen datos diarios de precipitación, evaporación, temperatura y datos de almacenamientos. El modelo de datos ArcHydro permite organizar los datos asociados a los recursos hídricos en una estructura estándar, integrando series de tiempo y datos espaciales en un sólo depósito de datos (base de datos geográfica), con lo cual se reducen los problemas asociados a la información. (Álvarez, 2010)

El Instituto Mexicano de Tecnología del Agua también cuenta con un Almacén de Datos que contienen datos asociados a los recursos hídricos.

En el campo del transporte de cargas y pasajeros se utilizan Almacenes de Datos para almacenar y acceder a meses o años de datos de clientes y sistemas de reservas para realizar actividades de marketing, planeamiento de capacidad, monitoreo de ganancias, proyecciones y análisis de ventas y costos, programas de calidad y servicio a clientes. En este caso aparecen empresas de la envergadura de Cornrail, Union Pacific, Norfolk Southern, American President Lines, Delta, Lufthansa, QANTAS, British Airways, Air France, American Airlines, Canadian Airlines y SNFC.

Las organizaciones relacionadas con los bancos, finanzas, el gobierno, las telecomunicaciones y la salud de forma general son los que más implementan Almacenes de Datos.

Cuba a pesar de los factores económicos que no son muy favorables para experimentar con la aplicación de las nuevas tecnologías ha incursionado también en el desarrollo de los Almacenes de Datos. Un ejemplo es la Corporación CIMEX la cual implementó un Almacén que centra su atención en la actividad del comercio con el objetivo de apoyar a la toma de decisiones en dicha entidad.

La Empresa de Proyectos de Arquitectura e Ingeniería (EMPAI) de Matanzas creada para dar respuesta a importantes inversiones del sector de la Construcción, de alto desempeño considerada como líder en el país en la esfera del Diseño y la Consultoría al igual que el Banco de Inversiones S.A, han trabajado en la implementación de Almacenes de Datos para ayudar a la toma de decisiones en estas entidades.

En la Universidad de las Ciencias Informáticas se llevan proyectos que implementan Almacenes de Datos para ayudar a la toma de decisiones de diferentes organizaciones como:

- La Oficina Nacional de Estadísticas (ONE): Entidad creada para garantizar la producción de estadísticas de calidad a través del Sistema Estadístico Nacional, ejerciendo una adecuada dirección, ejecución y control de la captación de las cifras económicas y sociales, así como su difusión de acuerdo con los requerimientos de la economía y las demás necesidades del país en información estadística.
- Universidad de las Ciencias Informáticas: El proyecto surgió como una necesidad de la UCI en cuanto a la disposición de información para los directivos, y poder tomar decisiones precisas ante cada situación particular en las distintas esferas de la universidad.

A partir del estudio del estado actual de los Almacenes de Datos se determinó que algunos de los sistemas encontrados se asemejan en parte al que se pretende desarrollar, sin embargo cada proceso de desarrollo de Almacenes de Datos es diferente debido a la necesidades de las organizaciones y muchos de estos

sistemas no presentan los indicadores que se desean analizar, por lo que ninguna de estas soluciones son factibles, y se decide implementar una solución que cubra las necesidades de la Dirección General Infraestructura y Servicios.

Conclusiones del capítulo

A partir del análisis del estado del arte y de la importancia de la utilización de los Almacenes de Datos para ayudar en la toma de decisiones, se pudo establecer que ninguna de las soluciones estudiadas cubre las necesidades de la Dirección General Infraestructura y Servicios. Se decide implementar una solución que facilite la integridad y disponibilidad de la información y de respuesta al problema existente, para ello se obtuvieron los resultados siguientes:

1. Se seleccionó la metodología adecuada para resolver la problemática.
2. La solución se diseñará según el esquema constelación de hechos.
3. Se utilizará el modo de almacenamiento ROLAP.
4. Se seleccionó como herramienta de modelado Visual Paradigm 6.4, como SGBD PostgresSQL en su versión 9.1, para los procesos de BI y ETL se trabajará con las herramientas Pentaho Data Integration 3.2.0, Data Cleaner 1.5.4, Pentaho Metadata Editor 3.7.0, BI-Server 3.6, Schema Workbench 3.2.1, Pentaho Report Designer 3.7.0.

CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS INFRAESTRUCTURA Y SERVICIOS.

Introducción

En este capítulo se definen los requerimientos del sistema, se diseña el diagrama de CUS, los subsistemas de integración y visualización. Se presenta la descripción de la arquitectura, ya que es uno de los desarrollos más importantes dentro de la construcción del software permitiendo representar la estructura del sistema, además se deja plasmado el diseño del modelo de datos, la naturaleza de los datos, y se detalla el análisis de los datos fuentes. De manera general, aborda el resultado de las fases de Estudio Preliminar, Requerimientos, Arquitectura y Diseño definidas por la metodología seleccionada para el desarrollo del Mercado de Datos Infraestructura y Servicios.

2.1 Evaluación de las áreas de la organización

La AP se encuentra dividida en diferentes Direcciones, en este caso se estará analizando la Dirección General Infraestructura y Servicios, la cual cuenta con las siguientes Direcciones Provinciales:

Vivienda: Controla el cumplimiento de la Ley General de la Vivienda y demás Disposiciones Legales en cumplimiento de la Política del Estado y del Gobierno, además de las decisiones que adoptan las respectivas Asambleas del Poder Popular y la Administración Provincial. También responde por el funcionamiento de la Dirección así como la tramitación de las quejas que se generen al nivel provincial.

Comercio: Dirige y controla la aplicación de la política, normas y procedimientos para la circulación, distribución, y comercialización de los productos alimenticios, no alimenticios y otros bienes de consumo, así como la prestación de servicios personales y técnicos, además de los asociados a los soportes tecnológicos del comercio de forma mayorista.

Recursos Hidráulicos: Dirige, supervisa y ejecuta en el territorio, la política nacional de los recursos hídricos e hidráulicos, además de exigir y velar por la protección, cuidado y conservación del patrimonio de la infraestructura hidráulica del territorio.

Transporte: Organiza y controla el cumplimiento de las regulaciones estatales, normas técnicas y disposiciones del Ministerio del Transporte, para los medios e instalaciones del transporte de subordinación municipal y provincial, los servicios de transportación de pasajeros y de cargas, servicios generales a vehículos ligeros.

La información estadística de la Dirección General Infraestructura y Servicios es registrada en los modelos que se manejan en cada una de las Direcciones Provinciales y contienen los indicadores referentes a los procesos que se llevan a cabo. Esta información es analizada en diferentes períodos de tiempo (diaria, semanal, mensual, trimestral, semestral, anual) realizando la mayoría de los cortes de información por municipios.

Partiendo de la descripción del negocio para lograr el buen desarrollo de la solución, es necesario establecer las posibles Áreas de Análisis.

2.2 Áreas de Análisis

Las áreas de análisis representan una agrupación de información según su propósito, aunque el criterio depende de las necesidades de la institución o empresa donde se aplica el sistema. Permite restringir el número de usuarios que acceden a los datos. Las áreas de análisis identificadas en este trabajo son cuatro:

1. Análisis de los indicadores de vivienda.
2. Análisis de los indicadores de transporte.
3. Análisis de los indicadores de recursos hidráulicos.
4. Análisis de los indicadores de comercio.

El análisis y consulta de la información que se maneja en la Dirección General y en cada una de las direcciones que integran a dicha dirección, no debe estar al

alcance de cualquier persona. Por tanto se requieren de niveles de acceso restringidos para garantizar su seguridad y rendimiento.

Roles y permisos

Un rol representa el papel que pone en práctica una persona al interactuar con el sistema en función de los permisos que tiene al utilizar el producto. En la Dirección General Infraestructura y Servicios se definieron seis roles y los permisos de acceso al sistema con el objetivo de llevar un control de los datos. A continuación se nombrarán los mismos:

InfraestructuraServicios: Tiene permiso de consulta sobre los datos del Mercado de Datos.

Comercio: Tiene permiso de consulta al Área de Análisis de los indicadores de Comercio.

Transporte: Tiene permiso de consulta al Área de Análisis de los indicadores de Transporte.

Vivienda: Tiene permiso de consulta al Área de Análisis de los indicadores de Vivienda.

RHidraulicos: Tiene permiso de consulta al Área de Análisis de los indicadores de Recursos Hidráulicos.

Administrador: Permiso de consulta y actualización sobre los datos del Mercado de Datos. Tiene acceso total a todas las Áreas de Análisis. Es el encargado de gestionar todo lo relacionado a los usuarios, los roles del sistema y los reportes.

Reglas del Negocio

Las reglas del negocio describen las políticas, normas, operaciones, definiciones y restricciones presentes en una organización, y son de vital importancia para alcanzar los objetivos misionales, pues actúan como un medio por el cual la estrategia es implementada. Especifican en un nivel adecuado de detalle lo que

una organización debe hacer. Las reglas del negocio deben ser expresadas en lenguaje natural y orientadas al negocio.

En el análisis se identificaron 54 reglas del negocio que se encuentran detalladas en el artefacto de Reglas del Negocio (CFRA-Government-0116_RNeg) y dentro de las cuales se encuentran las siguientes:

1. Una vez cargados los datos en el Mercado, no pueden existir campos nulos.
2. Total de vehículos activos= Sumatoria de la cantidad de vehículos activos.
3. Total de vehículos inactivos= Sumatoria de la cantidad de vehículos paralizados por motor, paralizados por batería, paralizados por neumáticos.

Después de definidas las reglas del negocio es importante determinar cuáles son los requisitos para el desarrollo de la solución.

Requisitos

La IEEE Standard Glossary of Software Engineering Terminology define un requerimiento como condición o capacidad que necesita un usuario para resolver un problema o lograr un objetivo. En los sistemas basados en Almacenes de Datos suelen especificarse tres tipos de requisitos: requisitos de información, requisitos funcionales y requisitos no funcionales.

Requisitos de Información

Los requisitos de información debido a que son las principales necesidades de los usuarios deben estar disponibles para estos en todo momento con el objetivo de brindar una mejora en el proceso de toma de decisiones. En este trabajo los requisitos de información se agruparon por las diferentes áreas de análisis.

Se identificaron 83 requisitos informativos que se encuentran descritos de forma detallada en el artefacto de Especificación de Requisitos de Software (CFRA-Government-0113_ERS), dentro de los cuales se encuentran los siguientes:

1- Obtener de los modelos indicadores de Recursos hidráulicos el tiempo de bombeo, el caudal de bombeo, el volumen bombeado y el volumen que se

acumula durante el bombeo diario, por estación de bombeo y por período.

2- Obtener de los modelos indicadores de Transporte la cantidad de vehículos total, la cantidad de activos, la cantidad de inactivos, el Coeficiente de Disponibilidad Técnica (CDT), la cantidad de viajes, la cantidad por motivos de paralización, por municipio, tipo de vehículo y por período.
3- Obtener de los modelos indicadores de Vivienda el plan, el plan base, el real y el por ciento por programa, por municipio y por período.
4- Obtener de los modelos indicadores de Comercio la cantidad de consumidores menores de 2 años, la cantidad de consumidores de 2 a 6 años, la cantidad de consumidores de 7 a 13 años, la cantidad de consumidores de 14 a 64 años, la cantidad de consumidores mayores de 65 años, por municipio y por período.

Requisitos Funcionales

Los requisitos funcionales son las capacidades o condiciones que el sistema debe cumplir. Se orientan a las necesidades de información de los usuarios. Los requisitos funcionales quedan plasmados de forma detallada en el artefacto de Especificación de Requisitos de Software (CFRA-Government-0113_ERS), siendo los siguientes:

RF 1 Autenticar usuario: Permite el acceso a la aplicación y define las actividades según el rol.

RF 2 Mostrar gráficos: Permite mostrar el gráfico asociado al reporte seleccionado.

RF 3 Configurar vistas de análisis OLAP: Permite configurar las diferentes vistas de análisis por las que se muestra un reporte.

RF 4 Realizar la transformación y carga de los datos: Permite realizar la trasformación a los datos extraídos, para luego ser cargados en el Mercado de Datos.

RF 5 Realizar la extracción de los datos: Permite realizar la extracción de los datos de los archivos fuentes.

RF 6 Adicionar usuario: Permite adicionar usuarios al sistema.

RF 7 Eliminar usuario: Permite eliminar un usuario existente del sistema.

RF 8 Adicionar reporte: Permite adicionar nuevos reportes al sistema.

RF 9 Eliminar reporte: Permite eliminar un reporte existente en el sistema.

RF 10 Adicionar rol: Permite adicionar nuevos roles al sistema.

RF 11 Eliminar rol: Permite eliminar un rol existente en el sistema.

RF 12 Crear reportes Ad-Hoc: Permite al usuario elaborar sus propios reportes.

Requisitos No Funcionales

Los requisitos no funcionales son las cualidades o propiedades que el sistema debe cumplir. Se identificaron 21 requisitos no funcionales y se encuentran detallados en el artefacto de Especificación de Requisitos de Software (CFRA-Government-0113_ERS). Los requisitos identificados pueden clasificarse en requisitos de usabilidad, fiabilidad, eficiencia, soporte, restricciones de diseño, de interfaces, de licencia, entre otros.

Casos de Uso del Sistema

Los Casos de Uso del Sistema son las acciones que el sistema debe permitir. El mismo es una técnica para la captura de requisitos y los actores que intervienen en el sistema. Se identificaron 16 CUS, de ellos 8 son informativos:

CUS1: Administrar usuario.

CUS2: Administrar reporte.

CUS3: Autenticar usuario.

CUS4: Transformar y cargar datos.

CUS5: Extraer datos.

CUS6: Visualizar la información de vivienda.

CUS7: Visualizar la información de transporte.

CUS8: Visualizar la información de la utilización de las aguas.

CUS9: Visualizar la información de la lluvia.

CUS10: Visualizar la información de la situación de los embalses.

CUS11: Visualizar la información de la red mayorista.

CUS12: Visualizar la información de la red minorista.

CUS13: Visualizar la información de la inspección.

CUS14: Administrar rol.

CUS15: Modificar reporte.

CUS16: Crear reportes Ad-Hoc.

La especificación de cada CUS así como la descripción de cada actor del sistema se encuentran en el artefacto de Modelo del Sistema (CFRA-Government-MCUSist). Se define el siguiente diagrama de CUS compuesto por nueve actores y las actividades que realizan cada uno de ellos.

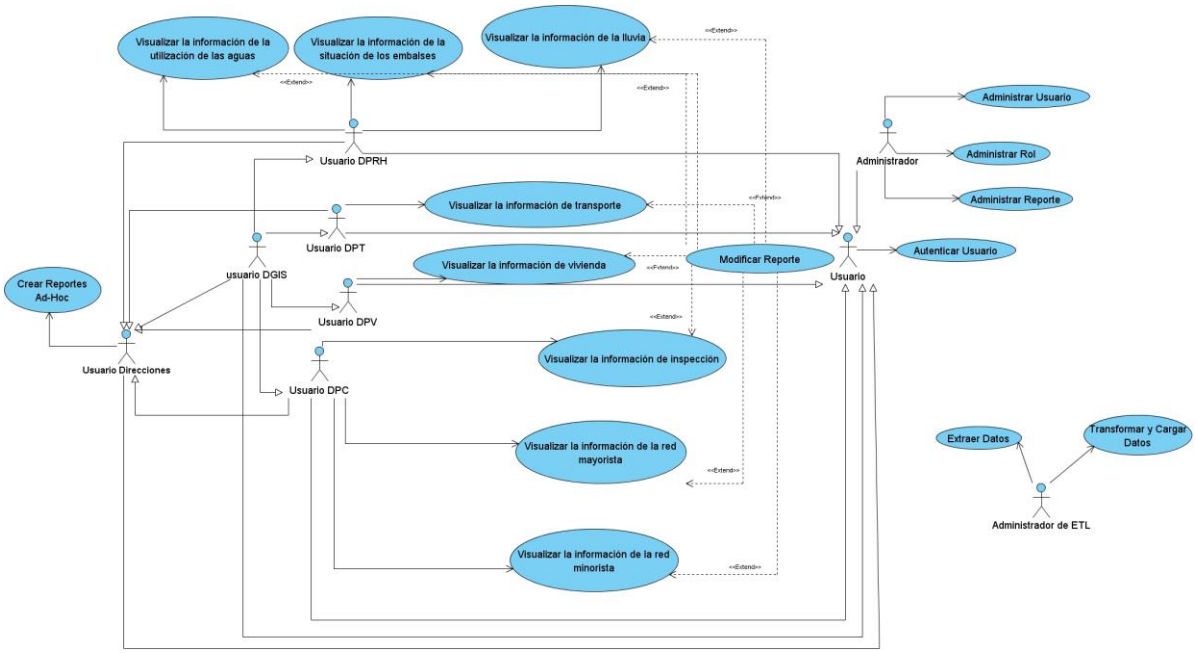


Ilustración 2 Diagrama de CUS

2.3 Arquitectura e interacción de los componentes del sistema

La arquitectura del software es uno de los aspectos más importantes que deben tenerse en cuenta en la construcción del mismo. A través de la arquitectura del software se puede representar la estructura que tendrá el sistema.

Para tener una visión general del sistema se explica a continuación la arquitectura de la solución propuesta, detallando cada uno de los subsistemas que lo

conforman. Se identificaron 3 subsistemas en los cuales el sistema está estructurado: Subsistema de integración, Subsistema de almacenamiento, Subsistema de visualización.

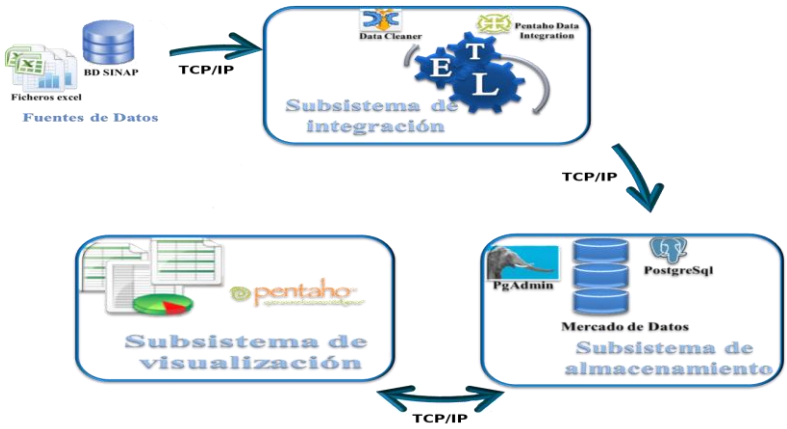


Ilustración 3 Representación de la arquitectura

El subsistema de integración es donde se agrupan los procesos que se encargan de llevar a cabo las tareas relacionadas con la extracción, integración, limpieza de los datos, carga y actualización del Mercado de Datos. Es donde se realizan todas las tareas desde que se extraen los datos de las fuentes hasta que se puebla el Mercado de Datos para su utilización. Para llevar a cabo estas tareas se utilizó la herramienta Pentaho Data Integration.

El subsistema de almacenamiento es el encargado de contener toda la información correspondiente al Mercado de Datos. Este estará compuesto por dimensiones y tablas de hechos que a su vez contendrán los datos que describirán un hecho. Para el almacenamiento de la información se hace uso del SGBD PostgreSQL y como aplicación gráfica el PgAdmin III.

El subsistema de visualización no es más que la capa de presentación, y tiene como finalidad mostrar los datos almacenados de forma útil a través de las distintas herramientas de Pentaho, es la capa con la cual interactúa el usuario final. Se comunica directamente con el servidor de cubos a través de consultas, las cuales

retornan la información requerida donde ésta es transformada y presentada para la visualización de los reportes y vistas de análisis.

2.4 Diseño del subsistema de almacenamiento

A continuación se explica el modelo de datos, compuesto por las dimensiones y hechos seleccionados y las medidas definidas. Para comprender todo lo relacionado con el proceso de diseño para la solución propuesta, se explica en detalle cada uno de los pasos seguidos para el Área de Análisis de Transporte.

Proceso de negocio

Después de haber dado una descripción del negocio de la Dirección General Infraestructura y Servicios, se pueden definir los siguientes procesos dentro de las actividades y funciones que se llevan a cabo en la Dirección de Transporte:

1. Balance municipal del transporte
2. Balance del transporte

Nivel de granularidad

La granularidad es el nivel de detalle que posee cada registro de una tabla de hechos. Además el gránulo determina las dimensiones básicas del sistema.

Proceso: Balance municipal del transporte

La información estadística semanal perteneciente a la Dirección de Transporte, registrada bajo la estructura de medios de transporte, en todos los municipios, captados en el modelo 002 del parte al MITRANS.

Proceso: Balance del transporte

La información estadística semanal perteneciente a la Dirección de Transporte, registrada bajo la estructura de vehículos, captados en el modelo 001 del parte al MITRANS.

Dimensiones

Después de seleccionado el grano del proceso de negocio, se pueden identificar las posibles dimensiones que podrán existir para el diseño. La solución propuesta

cuenta con 20 dimensiones, algunas de las cuales tienen entre sus características que presentan jerarquías entre sus atributos, a continuación se describen cada una de las dimensiones relacionadas con el Área de Análisis de Transporte:

dim_municipio: describe el universo de valores bajo los cuales puede clasificarse la información atendiendo al municipio.

dim_municipio		
+dim_municipio_id	**int4**	**Nullable = false**
municipio_codigo	char(4)	Nullable = false
municipio_nombre	varchar(26)	Nullable = false
municipio_descripcion	varchar(15)	Nullable = false

dim_temporal_dia: define una línea de tiempo para enmarcar la información almacenada. Representa jerárquicamente cuando fue captada la información.

dim_temporal_dia		
+dim_temporal_dia_id	**int4**	**Nullable = false**
anno_codigo	varchar(4)	Nullable = false
anno_numero	int4	Nullable = false
anno_nombre	char(4)	Nullable = false
semestre_codigo	varchar(6)	Nullable = false
semestre_numero	int4	Nullable = false
semestre_nombre	char(11)	Nullable = false
trimestre_codigo	varchar(6)	Nullable = false
trimestre_numero	int4	Nullable = false
trimestre_nombre	char(11)	Nullable = false
mes_codigo	char(6)	Nullable = false
mes_numero	int4	Nullable = false
mes_nombre	char(25)	Nullable = false
semana_codigo	varchar(4)	Nullable = false
semana_numero	int4	Nullable = false
semana_nombre	char(15)	Nullable = false
dia_codigo	varchar(2)	Nullable = false
dia_numero	int4	Nullable = false
dia_nombre	char(11)	Nullable = false

dim_vehiculo: describe el universo de valores bajo los cuales puede clasificarse la información atendiendo al vehículo.

dim_vehiculo		
+dim_vehiculo_id	**int4**	**Nullable = false**
vehiculo_tipo_codigo	char(3)	Nullable = false
vehiculo_tipo	varchar(20)	Nullable = false
vehiculo_marca_codigo	char(3)	Nullable = false
vehiculo_marca	varchar(20)	Nullable = false

dim_medio_transporte: describe el universo de valores bajo los cuales puede clasificarse la información atendiendo al medio de transporte.

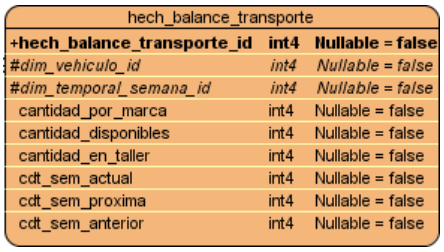
dim_medio_transporte		
+dim_medio_transporte_id	**int4**	**Nullable = false**
clasificacion_transporte_codigo	char(3)	Nullable = false
clasificacion_transporte	varchar(30)	Nullable = false

Medidas y tablas de hechos

Como último paso es necesario identificar las medidas que surgen de los procesos de negocios. La solución propuesta cuenta con 24 tablas de hechos y 148 medidas que serán analizadas para apoyar el proceso de toma de decisiones. A continuación se definirán las tablas de hechos para el Área de Análisis de

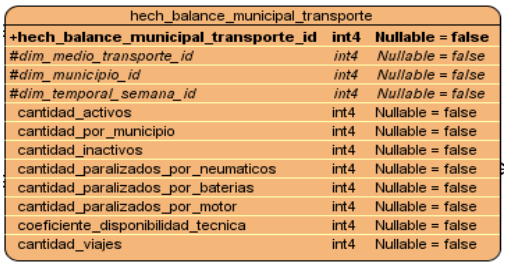
Transporte, que son aquellas que contendrán las medidas a través de las cuales se construirán los indicadores de análisis.

hech_balance_transporte: almacena las medidas relacionadas con los tipos de vehículos.



hech_balance_transporte		
+hech_balance_transporte_id	**int4**	**Nullable = false**
#dim_vehiculo_id	*int4*	*Nullable = false*
#dim_temporal_semana_id	*int4*	*Nullable = false*
cantidad_por_marca	int4	Nullable = false
cantidad_disponibles	int4	Nullable = false
cantidad_en_taller	int4	Nullable = false
cdt_sem_actual	int4	Nullable = false
cdt_sem_proxima	int4	Nullable = false
cdt_sem_anterior	int4	Nullable = false

hech_balance_municipal_transporte: almacena las medidas relacionadas con los medios de transporte por municipio.



hech_balance_municipal_transporte		
+hech_balance_municipal_transporte_id	**int4**	**Nullable = false**
#dim_medio_transporte_id	*int4*	*Nullable = false*
#dim_municipio_id	*int4*	*Nullable = false*
#dim_temporal_semana_id	*int4*	*Nullable = false*
cantidad_activos	int4	Nullable = false
cantidad_por_municipio	int4	Nullable = false
cantidad_inactivos	int4	Nullable = false
cantidad_paralizados_por_neumaticos	int4	Nullable = false
cantidad_paralizados_por_baterias	int4	Nullable = false
cantidad_paralizados_por_motor	int4	Nullable = false
coeficiente_disponibilidad_tecnica	int4	Nullable = false
cantidad_viajes	int4	Nullable = false

En el artefacto Especificación del Modelo Dimensional (CFRA-Government-Especificación del Modelo Dimensional) se encuentran declarados los gránulos, dimensiones, medidas, y tablas de hechos de cada una de las Áreas de Análisis que componen el Mercado de Datos.

2.5 Modelo dimensional

Luego de haber definido las dimensiones, las medidas, la granularidad, se procede a la estructuración del modelo dimensional. En este trabajo se utilizó el esquema constelación de hechos pues contribuye a la reutilización de dimensiones debido a que una misma dimensión puede utilizarse para varias tablas de hechos.

La idea fundamental del modelo multidimensional es que los datos de negocio pueden ser representados como un tipo de cubo de datos. En los cubos cada celda contiene un valor y las aristas del cubo definen dimensiones naturales de análisis.

En la solución propuesta existirán 24 tablas centrales las cuales estarán relacionadas con las dimensiones descritas. Con este fin cada dimensión posee una llave primaria que es la encargada de mantener la integridad referencial entre

ellas y la tabla de hechos. Esta llave primaria no posee ningún tipo de significado dentro del negocio, solo es un número que garantiza las uniones.

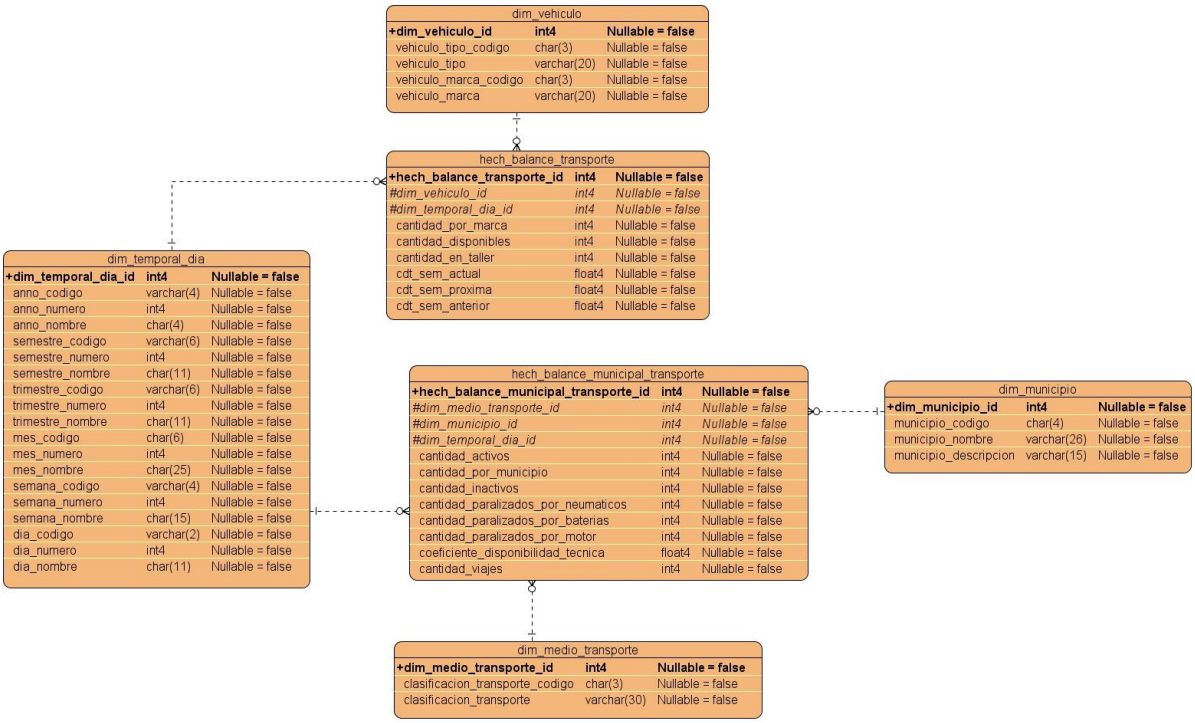


Ilustración 4 Modelo de Datos Dimensional Área de Análisis Transporte

En el artefacto Especificación del Modelo Dimensional (CFRA-Government-Especificación del Modelo Dimensional) se encuentra detallado todo lo relacionado con los modelos de datos dimensionales para la solución propuesta.

2.6 Diseño del subsistema de integración

Perfilado de datos

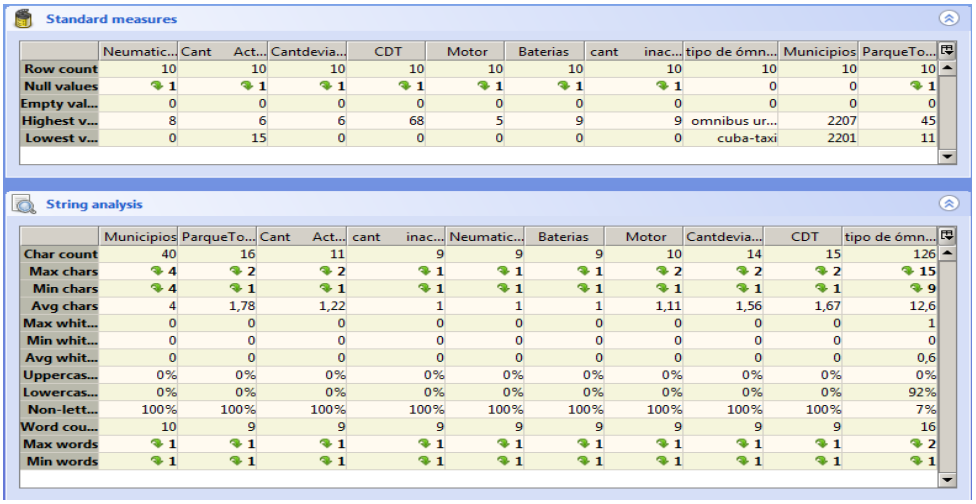
Antes de comenzar con el diseño de los procesos de ETL es necesario realizar el perfilado de datos con el objetivo de analizar los datos de los sistemas fuentes para comprender su contenido, estructura, calidad y dependencias.

Este paso es de vital importancia a la hora de hacer un análisis de los datos de origen pues muchas veces no se sabe que preguntar, ni donde pueden estar algunos problemas acerca de los mismos. De esta manera el perfilado brinda un análisis exacto de la estructura y modelo de datos a tratar, es por esto que se utilizó

para este proceso la herramienta Data Cleaner, la cual brinda hechos sobre los que construir el diseño de los procesos ETL. Esta herramienta permitió generar reportes de los Estándares de medidas, Análisis de cadenas, también permite generar otros reportes pero estos fueron los principales que permitieron realizar un efectivo análisis de los datos extraídos.

El resultado de este proceso se encuentra detallado en el artefacto Modelo de Integración de datos (CFRA-Government-Modelo_IntegraciónDatos). Durante la ejecución de este paso del análisis de los sistemas fuentes, se logró detectar un conjunto de errores que pueden tener cierta repercusión en el resultado final y permitieron llevar a cabo las transformaciones que se realizaron a través de los procesos ETL.

En los Estándares de medidas se obtuvo el resultado de todas las filas, valores nulos y vacíos, así como el mejor y más bajo valor de la fuente. Con el reporte del Análisis de cadenas se adquirió toda la información de la cantidad, mínima, máxima y promedio de caracteres, los espacios en blanco, los caracteres no escritos, entre otros parámetros, los cuales son mostrados a continuación.



Standard measures

	Neumatic...	Cant	Act...	Cantdevia...	CDT	Motor	Baterias	cant	inac...	tipo de ómn...	Municipios	ParqueTo...
Row count	10	10	10	10	10	10	10	10	10	10	10	10
Null values	1	1	1	1	1	1	1	1	1	0	0	1
Empty val...	0	0	0	0	0	0	0	0	0	0	0	0
Highest v...	8	6	6	68	5	9	9	9	9	omnibus ur...	2207	45
Lowest v...	0	15	0	0	0	0	0	0	0	cuba-taxi	2201	11

String analysis

	Municipios	ParqueTo...	Cant	Act...	cant	inac...	Neumatic...	Baterias	Motor	Cantdevia...	CDT	tipo de ómn...
Char count	40	16	11	9	9	9	10	14	15	126		
Max chars	4	2	2	1	1	1	2	2	2	15		
Min chars	4	1	1	1	1	1	1	1	1	9		
Avg chars	4	1,78	1,22	1	1	1	1,11	1,56	1,67	12,6		
Max whit...	0	0	0	0	0	0	0	0	0	1		
Min whit...	0	0	0	0	0	0	0	0	0	0		
Avg whit...	0	0	0	0	0	0	0	0	0	0,6		
Uppercas...	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%		
Lowercas...	0%	0%	0%	0%	0%	0%	0%	0%	0%	92%		
Non-lett...	100%	100%	100%	100%	100%	100%	100%	100%	100%	7%		
Word cou...	10	9	9	9	9	9	9	9	9	16		
Max words	1	1	1	1	1	1	1	1	1	2		
Min words	1	1	1	1	1	1	1	1	1	1		

Ilustración 5 Perfilado de los Estándares de medidas y Análisis de cadenas del modelo Parte al MITRANS 002 de la Dirección de Transporte

Diseño de los procesos de ETL

Previo a comenzar a diseñar los procesos de extracción, transformación y carga que se realizaron a los datos, fue necesario establecer un estándar para organizar el trabajo. Para dar cumplimiento a esto, se determinó nombrar a todas las transformaciones y trabajos con el prefijo “fc_etl”. De igual forma, el nombre de cada paso en la transformación debe sugerir la acción que se realiza.

Los datos de origen para la aplicación de la Inteligencia de Negocios provienen de varias fuentes. Es importante tener en cuenta la fuente de datos y los destinos de los mismos, con el objetivo de identificar de donde provienen los datos y sus características. Una vez seleccionadas las tablas en las que se va a trabajar, es necesario definir el orden y la secuencia de las transformaciones para cada conjunto de datos. Es necesario señalar que todas las tablas de dimensiones deben ser cargadas antes que las tablas de hechos.

El propósito fundamental de este proceso es fusionar esos datos en un solo formato para el Mercado de Datos.

Diagrama de actividades para el proceso de ETL

El diseño de los procesos de ETL se realiza con la interrelación de las actividades que se llevan a cabo dentro del mismo, las de extracción, transformación y carga de los datos. La siguiente figura muestra el diagrama de actividades para el proceso de ETL y comienza con la extracción de los datos de las fuentes, en este caso la base de datos del Sistema Informativo de la Administración Provincial (SINAP). Se comprueba que la información que se va a cargar esté en un período de tiempo determinado para evitar tener información duplicada en el Mercado de Datos. Luego se transforman los datos que se van a cargar en las tablas de hechos. Se buscan los identificadores de las tablas de dimensiones que se relacionan con las tablas fuentes, se validan estos identificadores, en caso de que exista algún problema con la validación los datos se almacenan en la tabla de hecho error correspondiente a la información que se esté cargando, en caso

contrario los datos son cargados en la tabla de hechos y por último se actualiza la fecha en que se cargó la información en la tabla de metadatos. De forma general estos son todos los pasos que se definieron para poblar el Mercado de Datos Infraestructura y Servicios.

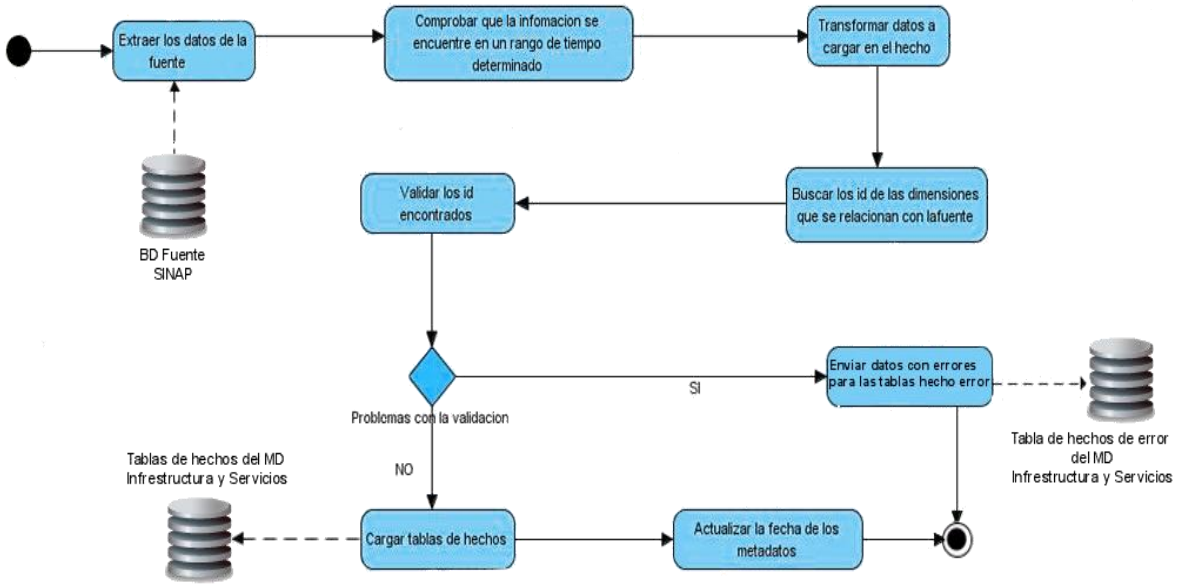


Ilustración 6 Proceso ETL para las cargas incrementales. Diagrama de actividades

Para las cargas históricas no fue necesario comprobar a que período de tiempo correspondían y los datos con errores se almacenaron en un fichero Excel.

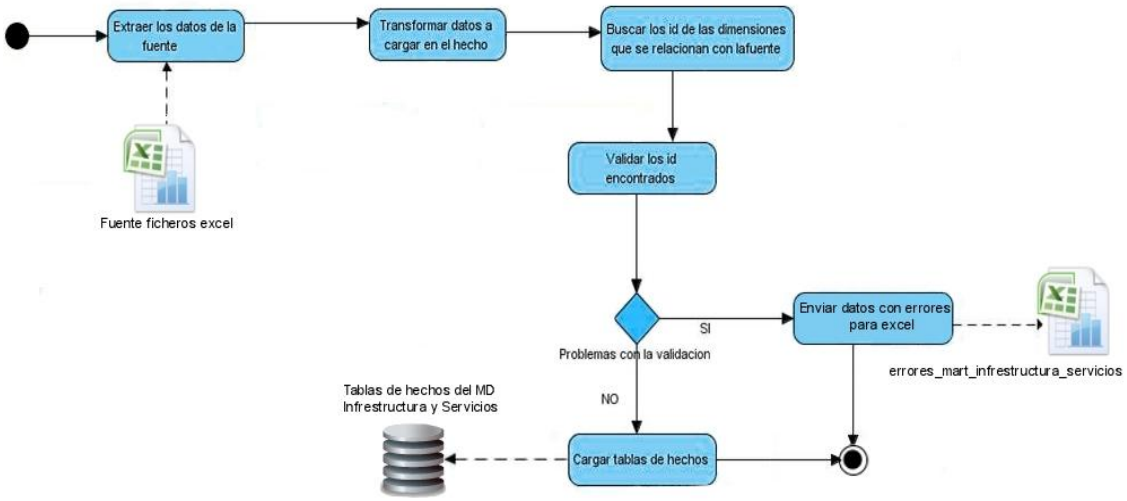


Ilustración 7 Proceso ETL para las cargas históricas. Diagrama de actividades

2.7 Diseño del subsistema de visualización

La presentación de la información a los usuarios finales del negocio es una de las partes más importantes de todo este proceso. Sería entonces la interfaz de publicación del BI-Server la que se utilizará en este sentido. De tal manera que el usuario que se encargará de elaborar las consultas y reportes tiene que poseer conocimientos básicos de cómo trabajar con esta herramienta.

A continuación se detallan los elementos que componen las estructuras de navegación de la información que será presentada en la capa de visualización del Mercado de Datos Infraestructura y Servicios; la misma contiene 4 Áreas de Análisis, 8 Libros de Trabajos y 83 reportes multidimensionales asociados a los Libros de Trabajos.

La siguiente imagen muestra el mapa de navegación en forma de árbol de secciones, niveles y contenidos relacionados con las Aéreas de Análisis, Libros de Trabajo y reportes que se identificaron.

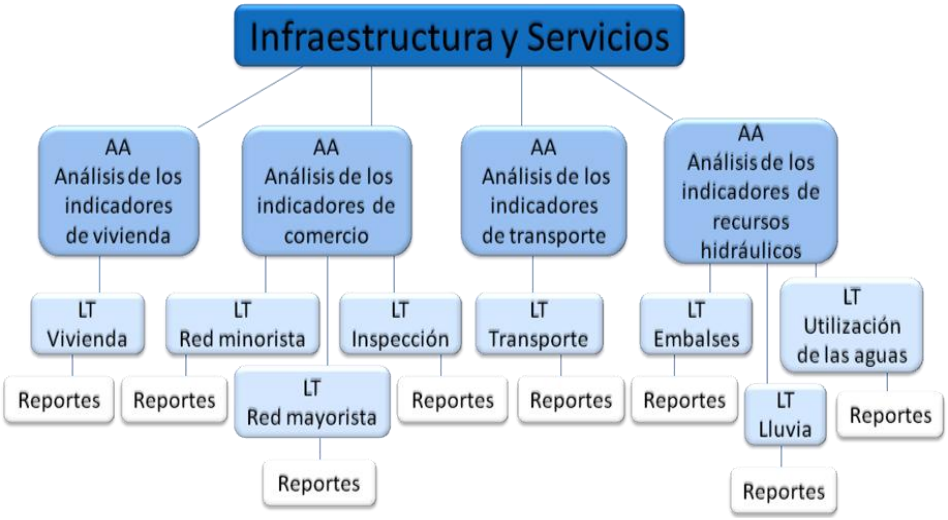


Ilustración 8 Arquitectura de Información

Descripción de las Áreas de Análisis

A.A Análisis de los indicadores de vivienda: Agrupa la información referente a todo lo relacionado con la vivienda.

A.A Análisis de los indicadores de comercio: Agrupa la información referente al comportamiento de la red de comercio mayorista y minorista, así como toda la información relacionada con la inspección en ese sector.

A.A Análisis de los indicadores de transporte: Esta área agrupa lo relacionado con el transporte tanto de carga como de pasajeros.

A.A Análisis de los indicadores de recursos hidráulicos: Agrupa la información referente al comportamiento de la lluvia, la situación de los embalses y la utilización de las aguas.

Descripción de los Libros de Trabajo

LT.1 Vivienda: Libro de trabajo contenido dentro del Área de Análisis de los indicadores de vivienda, contiene 21 reportes que permiten el análisis de datos correspondiente a este sector de la sociedad.

LT.2 Red minorista: Libro de trabajo contenido dentro del Área de Análisis de los indicadores de comercio, soportado por 34 reportes relacionados con esta área.

LT.3 Red mayorista: Libro de trabajo contenido dentro del Área de Análisis de los indicadores de comercio, contiene 3 reportes relacionados con esta área.

LT.4 Inspección: Libro de trabajo contenido dentro del Área de Análisis de los indicadores de comercio, contiene 2 reportes relacionados con esta área.

LT.5 Transporte: Libro de trabajo contenido dentro del Área de Análisis de los indicadores de transporte, incluye un conjunto de 10 reportes los cuales ayudan a la toma de decisiones ajustadas a los principales problemas que se presenten en esta área.

LT.6 Lluvia: Libro de trabajo contenido dentro del Área de Análisis de los indicadores de recursos hidráulicos, contiene 6 reportes que permiten el monitoreo de la lluvia caída en los municipios de la provincia.

LT.7 Embalses: Libro de trabajo contenido dentro del Área de Análisis de los indicadores de recursos hidráulicos, contiene 5 reportes que permiten el monitoreo de los principales embalses de la provincia.

LT.8 Utilización de las aguas: Libro de trabajo contenido dentro del Área de Análisis de los indicadores de recursos hidráulicos, contiene 2 reportes que permiten el monitoreo de la utilización de las aguas por los usuarios de la provincia.

Conclusiones de capítulo

En este capítulo se detallaron los aspectos principales dentro del proceso de análisis y diseño de la solución. Concluyéndose los siguientes resultados:

1. Se definieron las áreas de análisis, obteniendo así las áreas sobre las cuales se realizarán los análisis de la información.
2. Se determinaron las reglas del negocio.
3. Quedaron definidos los requisitos funcionales, no funcionales, y de información, para dar respuesta a las necesidades de los usuarios.
4. Fueron descritos los distintos componentes para la arquitectura que se ha propuesto para el Mercado de Datos Infraestructura y Servicios.
5. Se definieron 24 tablas de hechos y 20 tablas de dimensiones.
6. Se diseñaron los subsistemas de almacenamiento, integración y visualización.

CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS INFRAESTRUCTURA Y SERVICIOS.

Introducción

En este capítulo se implementan los procesos de extracción, transformación y finalmente la carga al Mercado de Datos, y también la implementación del subsistema de visualización, mostrándose las potencialidades de las herramientas seleccionadas en el Capítulo 1 para el desarrollo del Mercado de Datos destinado a la Dirección General Infraestructura y Servicios.

Una vez concluida la implementación del subsistema de almacenamiento del Mercado de Datos, se procede a la carga de los datos mediante los procesos de ETL hacia su destino según corresponda y finalmente se implementa la automatización de todo el proceso.

3.1 Implementación del subsistema de almacenamiento

Para la implementación del subsistema de almacenamiento se necesita la utilización de un gestor de base de datos, en este caso el utilizado fue el SGBD PostgreSQL utilizando el diseño elaborado y terminado anteriormente, lo que condujo a crear todas las tablas con sus respectivos atributos y relaciones.

El desarrollo exitoso de un modelo físico es un aspecto importante dentro de la construcción de un Almacén de Datos. El punto de partida para llevar a cabo este modelo es el modelo lógico, el cual no es más que el modelo dimensional elaborado en el capítulo anterior.

A continuación se describe todo lo relacionado con la implementación del modelo de datos físico en el Mercado de Datos Infraestructura y Servicios por lo que se manifestará la estructura de la base de datos, y se describirán los objetos que la conforman como son: esquemas, tablas, secuencias.

Esquemas y tablas

Los esquemas y las tablas permiten brindar una idea general de la base de datos, así como analizar y profundizar los contenidos básicos de la misma. Para la solución propuesta se cuenta con 44 tablas identificadas, para una buena organización de las mismas se agruparon por esquemas, las dimensiones comunes a los demás Mercados de Datos del Almacén de Datos para la AP de Artemisa están contenidas dentro del esquema “dimensiones”. Las dimensiones específicas del Mercado así como las tablas de hechos están agrupadas en el esquema “mart_infraestructura_servicios”, además las tablas de errores del Mercado de Datos se almacenarán en el esquema “errores_mart_infraestructura_servicios”. La información perteneciente a la última actualización de los hechos se encontrará en el esquema “metadatos”.

Restricciones y secuencias

Las restricciones son limitaciones que se desean imponer en el sistema, de forma que sea imposible almacenar datos incorrectos. Estas proveen un método de implementar reglas en la base de datos y restringen los datos que pueden ser almacenados en las tablas.

Las secuencias son atributos que se incrementan secuencialmente durante el ingreso de los datos. En el presente trabajo se definieron aproximadamente 41 llaves primarias y 41 secuencias.

3.2 Optimización

Un aspecto importante a tener en cuenta en el desarrollo de la solución propuesta es la optimización. Con respecto al Mercado de Datos se utilizaron índices con el objetivo de mejorar el tiempo de respuesta de las consultas que se realicen sobre el mismo y ganar en eficiencia.

Un índice no es más que una estructura física que permite un tipo de acceso alternativo al secuencial. Es creado a partir de una o varias columnas de una tabla,

y, por lo general, es construido en forma de árbol balanceado (B-Tree). (Sierra, 2009) Los índices son campos que permiten realizar búsquedas con mayor rapidez.

Para la solución se implementó como indexado el que trae por defecto el gestor PostgreSQL, que es el método Árboles-B (B-tree). De forma general se crearon índices para todos los campos llaves foráneas de las tablas de hechos, ya que los índices de las llaves primarias PostgreSQL los crea por defecto. De esta forma se minimizan las demoras al realizar JOINS entre las dimensiones y los hechos.

Se indexaron también aquellos campos por los que se realizan más búsquedas en el Mercado de Datos, para de esta forma optimizar los tiempos de respuesta de las consultas.

3.3 Implementación de los procesos de ETL

Implementación de las transformaciones

La transformación de los datos se realizó de acuerdo a las reglas que se definieron en el negocio. Estos procesos ETL para la solución propuesta no tuvieron la amplitud que normalmente requieren, debido a que la información histórica con que se cuenta es muy poca y además se encuentra con cierto grado de limpieza y calidad. Esto ocurre como consecuencia de los pasos previos que se realizan antes de llenar los Excel.

Los procesos de ETL que más peso tuvieron durante la implementación fueron los de extracción de los datos de las fuentes y la carga hacia el Mercado de Datos. Para llevar a cabo este proceso se trabajó con la herramienta Pentaho Data Integration.

Para poblar el Mercado de Datos se realizaron dos tipos de cargas: cargas históricas a partir de la información en ficheros Excel con la cual ya contaba la DGIS y cargas incrementales teniendo como fuente de datos la base de datos del SINAP, se realizó una transformación para cada tabla de hechos.

Debido a la extensión de la solución, a continuación solo se explicarán las transformaciones realizadas para cargar una de las tablas de hechos del Área de Análisis de Transporte.

La transformación comienza extrayendo los datos del fichero fuente (tablas para el diseño.xls). Luego de tener los datos se procede a la selección de los mismos y se renombran con el componente Seleccionar/Renombrar valores, con el objetivo de renombrar los valores del documento Excel para que coincida con los que están en el Mercado. El próximo paso es partir el campo fecha, luego se añaden constantes para capturar el día, el mes y el año, el componente Valor Javascript Modificado se utiliza para almacenar la fecha en cada constante, además de llenar los campos nulos con cero según las reglas de negocio identificadas.

El otro paso es realizar las búsquedas en la dim_vehiculo y dim_temporal_dia para poder relacionar los datos del flujo con los atributos de la tabla de hechos hech_balance_transporte. El componente de filtrado de filas permite filtrar los datos, en este caso se filtró por el identificador de la tabla dim_vehiculo y el identificador de la tabla dim_temporal_dia, con el objetivo de darle tratamientos diferentes a la información. En caso de que los identificadores sean nulos los datos son enviados a un fichero Excel error_mart_infraestructura.xls en caso contrario los datos son almacenados satisfactoriamente en la tabla de hechos.

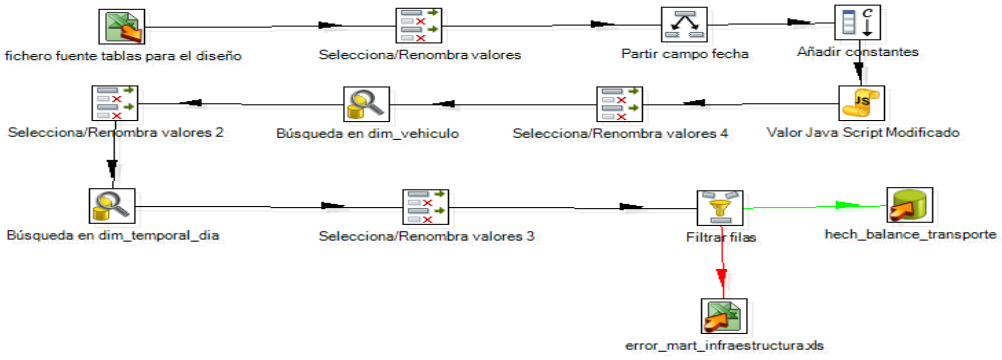


Ilustración 9 Transformación para realizar la carga a la tabla de hechos hech_balance_transporte

Este mismo procedimiento se siguió para realizar las cargas de la base de datos fuente, en este caso se toma como entrada la tabla dpartemitranstransporte, además se utilizan los componentes partir campos para capturar la fecha. Al igual que en la transformación anterior los datos son filtrados por los identificadores dim_vehiculo y dim_temporal_dia, la información correcta se almacenará en el Mercado, esto se hará a través del componente Insertar/Actualizar y si los datos son erróneos se almacenarán en la tabla de hecho error hech_error_balance_transporte.

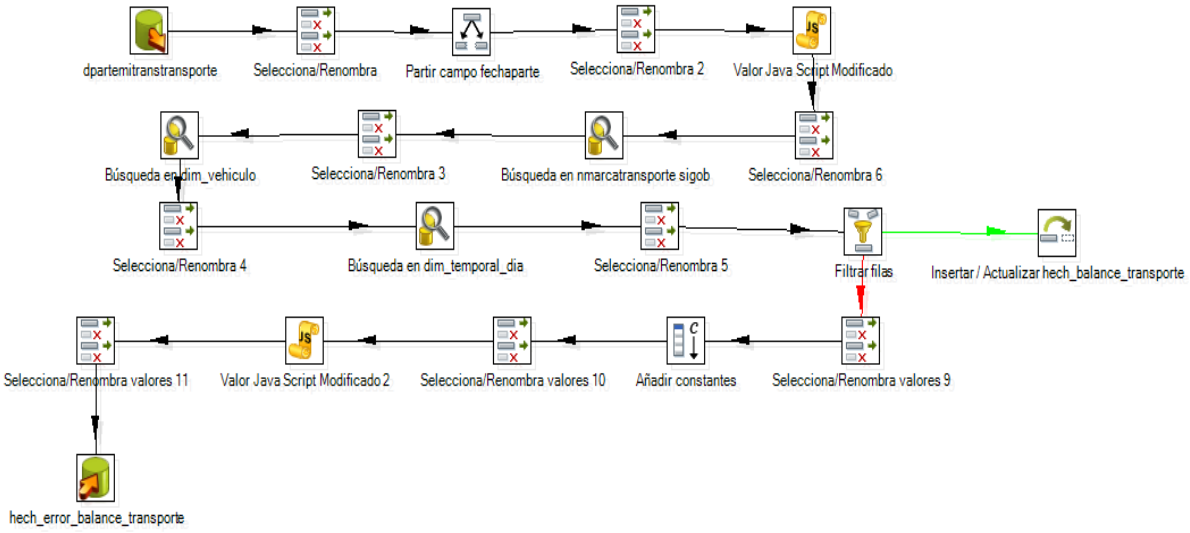


Ilustración 10 Transformación para realizar la carga a la tabla de hechos hech_balance_transporte

Implementación de los Trabajos

Una vez que se han realizado todas las transformaciones necesarias, es preciso organizar el proceso de carga al Mercado de Datos, para esto se aplican los Trabajos, los cuales están pensados para controlar la ejecución e interacción de las transformaciones y se le define la frecuencia con que serán cargados los datos al Mercado. De forma general se realizó un trabajo por cada tabla de hechos y uno general que ejecute a los demás en dependencia de la frecuencia de carga (diario, semanal, mensual) de la información correspondiente a cada dirección.

Teniendo en cuenta la frecuencia de actualización de la base de datos para el SINAP, se propone la automatización de los procesos de ETL.

El objetivo principal de esta automatización es garantizar el mantenimiento o actualización de los datos, o sea, añadir al Mercado de Datos aquellos datos nuevos que fueron generados después de la última carga. La misma se llevó a cabo mediante la programación de los trabajos diseñados, haciendo uso del Pentaho Data Integration.

Este proceso de automatización será descrito a continuación para la Dirección de Transporte:

En el caso de esta Dirección la información es cargada semanalmente (todos los martes) a las 2:00 de la mañana, y el orden en que se van a ejecutar los trabajos es de la manera siguiente: primero un trabajo que actualice la tabla de hechos hech_balance_municipal_transporte y luego uno que actualice la tabla de hechos hech_balance_transporte, mediante el trabajo general fc_etl_trabajo_cargas_transporte.

Ilustración 11 Trabajo para actualizar las tablas de hechos del Área de Análisis de Transporte

Los dos primeros trabajos mencionados anteriormente a su vez llaman a un trabajo que es el encargado de cargar la información a la tabla de hechos, por ejemplo en el caso de fc_etl_trabajo_hech_balance_transporte las transformaciones se ejecutan en el siguiente orden: primero se obtiene la fuente de la que se va a cargar (fc_etl_obtener_xml_dpartemitranstransporte). El siguiente paso es obtener la fecha del servidor donde se encuentra la tabla fuente y la fecha de la última vez que cargó información al Mercado, esto último se hace consultando la tabla de metadatos del Mercado de Datos (fc_etl_obtener_fechas_dpartemitranstransporte).

Luego se llama a la transformación para llenar la tabla de hechos que fue explicada en el acápite anterior (fc_etl_hech_balance_transporte) y por último se actualiza la tabla de metadatos (fc_etl_actualizar_metadatos).

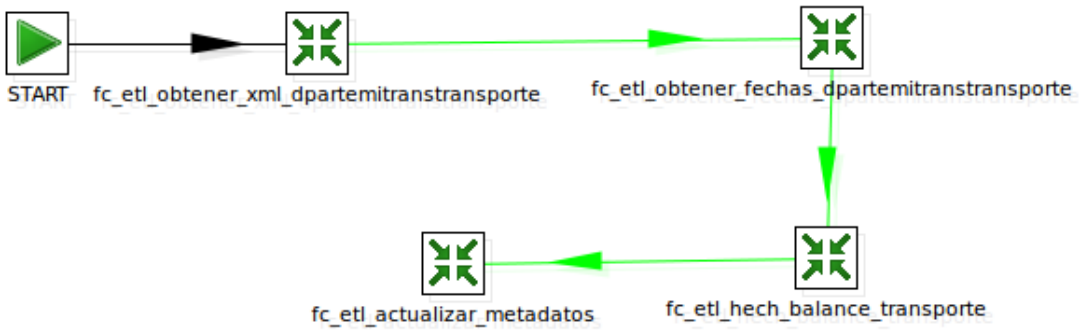


Ilustración 12 Trabajo para actualizar la tabla de hechos hech_balance_transporte

Este fue el procedimiento seguido para cada conjunto de transformaciones y trabajos realizados con el objetivo de poblar el Mercado de Datos.

3.4 Implementación del subsistema de visualización

Cubos de datos

En el sistema se definieron 25 cubos OLAP, que abarcarán todas las dimensiones descritas anteriormente. Dichos cubos se nombrarán de acuerdo a los procesos de negocio con los cuales estén relacionados y contendrán las medidas necesarias para regir el hecho numérico en esta estructura y según los objetivos estratégicos a cumplir. Para la implementación del subsistema de visualización de los datos se siguieron las reglas del negocio.

Los cubos de datos se desarrollaron mediante el uso de la herramienta Schema Workbench. El tipo de almacenamiento definido para estos cubos es el ROLAP.

El primer paso para la creación de los cubos es la definición de los mismos y donde se deciden cuáles son los atributos que son necesarios para realizar el análisis. Como segundo paso se especifica la estructura de las dimensiones diseñando los campos calculables, las medidas, junto con la creación de todas las tablas relacionales requeridas.

Para poder realizar el análisis de los cubos OLAP estos deben publicarse dentro de la plataforma Pentaho. A continuación se presenta la estructura general de los cubos utilizados dentro de la solución propuesta.

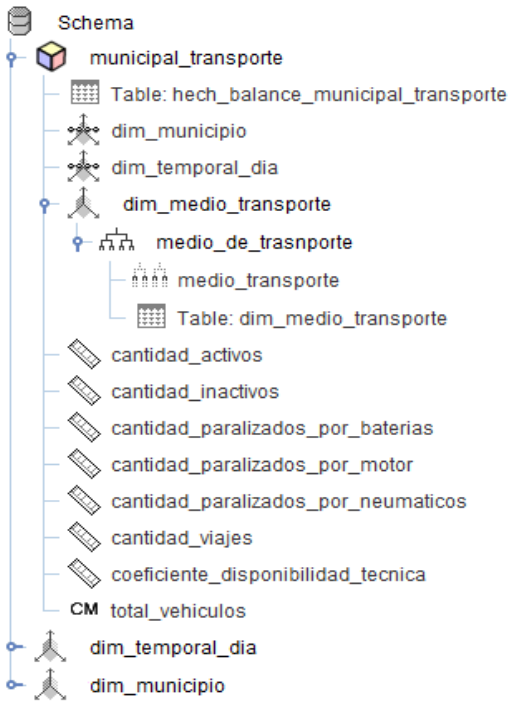


Ilustración 13 Representación del cubo municipal_transporte

En el caso de la figura anterior el cubo representado contiene un conjunto de dimensiones que se encuentran en el primer nivel, también la tabla de hechos de la misma (hech_balance_municipal_transporte), las medidas que se desean registrar y por último un miembro calculable.

Vistas de análisis

La solución propuesta cuenta con 48 vistas de análisis, con el objetivo de satisfacer las necesidades de información de la Dirección General Infraestructura y Servicios.

El Pentaho BI-Server permite mostrar el resultado del análisis realizado, pues a través de la misma es posible ver los reportes y analizar la información mediante tablas de datos o gráficas de diversos tipos (barras, pastel, líneas); permitiendo

además, desplegar el cubo de información, y modificar las vistas de análisis o crear unas totalmente nuevas. Por otra parte, esta interfaz brinda la posibilidad de cambiar el tipo de gráfica, imprimir el reporte o salvarlo en un archivo de formato PDF o XLS.

Reportes

En la implementación de la solución se generaron 23 reportes útiles para la Dirección General Infraestructura y Servicios, a partir de los datos registrados.

Para la definición de los reportes se utilizó la herramienta Report Designer de Pentaho. Esta herramienta permite consultar los datos de distintas fuentes y ponerlos a disposición de los usuarios en diferentes formatos (HTML, PDF, Microsoft Excel, y texto plano). Las herramientas de Pentaho permiten distribuir los reportes a los usuarios interesados, así como también publicarlos para que los usuarios puedan observar la información que necesitan.

Además de los reportes definidos en el Report Designer y publicados en la plataforma de Pentaho es posible que el usuario construya sus propios reportes a partir de la información de los metadatos definidos en el sistema con el uso de la herramienta Pentaho Metadata Editor.

2.5 Guía de Implantación

En el proceso de implantación del Mercado de Datos se hace necesario seguir una guía coherente y ordenada de pasos que se relacionan a continuación:

- Debe de estar instalado PostgreSQL 9.1 como gestor de base de datos.
- Debe estar instalada una herramienta de administración de base de datos, la cual ya fue definida: PgAdmin III versión 1.14.
- Se debe crear la nueva base de datos utilizando la herramienta de administración.
- Se deben crear los roles.

- Se debe crear un esquema para las dimensiones comunes, un esquema para las tablas de hecho error, uno para los metadatos y otro que contenga las tablas propias del Mercado de Datos Infraestructura y Servicios.
- Se debe crear un esquema para las dimensiones comunes, un esquema para las tablas de hecho error, uno para los metadatos y otro que contenga las tablas propias del Mercado de Datos Infraestructura y Servicios.
- Se debe cargar el script de lenguaje de definición de datos (DDL) de las dimensiones y hechos.
- Se debe cargar el script de lenguaje de control de datos (DCL) para asignar los permisos a los roles.

Conclusiones del capítulo

Durante el desarrollo de este capítulo se describió la implementación de los subsistemas de almacenamiento, integración y visualización, así como los pasos para la implantación del sistema. Concluyéndose que:

1. Se detalló la estructura física del Mercado de Datos, logrando poblarlo satisfactoriamente.
2. Se modeló el esquema dimensional con sus respectivos cubos OLAP.
3. Se implementaron los reportes correspondientes a las áreas de análisis que se identificaron previamente.
4. Se creó la guía de implantación para el Mercado de Datos.

CAPÍTULO 4: VALIDACIÓN DE LAS FUNCIONALIDADES DEL MERCADO DE DATOS INFRAESTRUCTURA Y SERVICIOS.

Introducción

Una vez desarrollado el análisis, diseño e implementación del Mercado de Datos para la Dirección General Infraestructura y Servicios, es necesario validar su funcionamiento y comprobar el éxito del mismo. Para cumplir este objetivo se realizan el Plan de pruebas y los Casos de Pruebas por cada Caso de Uso de información. Además se realizaron las pruebas de unidad e integración.

4.1 Pruebas

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

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

1. Verificar la interacción de componentes.
2. Verificar la integración adecuada de los componentes.
3. Verificar que todos los requisitos se han implementado correctamente.
4. Identificar y asegurar que los defectos encontrados se han corregido antes de entregar el software al cliente.

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

- **Prueba unitaria:** Es el proceso de probar los componentes individuales de la solución. El propósito es identificar diferencias entre la especificación de los artefactos y el comportamiento real de cada módulo.
- **Prueba de integración:** Es el proceso en el cual los componentes son

agregados para crear componentes más grandes. Es la prueba realizada para mostrar que aunque los componentes hayan pasado satisfactoriamente las pruebas de unidad, la integración de los componentes es incorrecta.

- **Prueba de sistema:** Se refiere al comportamiento del sistema integrado. La prueba de sistema se aplica generalmente para probar los requerimientos no funcionales de la solución.
- **Pruebas funcionales:** Aseguran el trabajo apropiado de los requisitos funcionales, incluyendo la navegación, entrada de datos, procesamiento y obtención de resultados.
- **Pruebas de aceptación:** El objetivo de las pruebas de aceptación es validar que un sistema cumpla con el funcionamiento esperado y permitir al usuario de dicho sistema que determine su aceptación, desde el punto de vista de su funcionalidad y rendimiento. Las pruebas de aceptación son definidas por el usuario del sistema y preparadas por el equipo de desarrollo, aunque la ejecución y aprobación final corresponden al usuario.

4.2 Plan de pruebas

El Plan de pruebas es un artefacto que describe el plan para probar las funcionalidades y características del Mercado de Datos Infraestructura y Servicios. Este documento está basado en los siguientes objetivos:

1. Establecer los roles y responsabilidades en esta etapa.
2. Identificar los escenarios de prueba.
3. Listar los requerimientos recomendados de prueba
4. Recomendar y describir las estrategias a ser empleadas.
5. Identificar los recursos requeridos para la realización de las pruebas.
6. Listar los elementos a entregar de las actividades de pruebas.

4.3 Pruebas funcionales

Las pruebas funcionales se realizaron a través de los Casos de Prueba basados en CU con el objetivo de validar si el comportamiento observado del software cumple o no con sus especificaciones.

El propósito de un Caso de Prueba es especificar una forma de probar el sistema, incluyendo las entradas para la validación, los resultados esperados y las condiciones bajo las que ha de probarse.

Para validar los requerimientos del sistema de este trabajo se le realizan Casos de Prueba a cada Caso de Uso informativo, con el objetivo de comprobar la disponibilidad de los perfiles de análisis y los indicadores a medir, así como también verificar el cumplimiento de los requisitos de información a través de los reportes candidatos.

Se diseñaron 8 Casos de Prueba basados en los Casos de Uso de Información, y se encuentran detallados en cada uno de los artefactos generados correspondientes a cada Caso de Uso (Caso de Prueba basado en CU).

4.4 Pruebas de integración

Para verificar el correcto funcionamiento de los procesos de ETL, se le realizaron pruebas a las transformaciones que se implementaron para cargar las tablas de hechos del Mercado de Datos para la Dirección General Infraestructura y Servicios de la Administración Provincial de Artemisa.

Las pruebas realizadas al subsistema de integración se encuentran documentadas en el artefacto Pruebas de Unidad e Integración, las cuales arrojaron resultados satisfactorios. La siguiente tabla muestra la prueba de integración realizada a la tabla de hechos balance transporte.

Prueba de: Integración		
Nombre Prueba: Transformación de la tabla de hechos balance transporte		
Estado: Satisfactoria	**Tipo**: Carga	**Última ejecución**: 21/05/2012
Ejecutado por: Sonia Fernández		**Verificado por**: Diana R. Prieto
Descripción: Para realizar la transformación de la tabla de hechos hech_balance_transporte los datos fueron extraídos de la fuente de datos (base de datos del SINAP), luego fueron transformados los datos para finalmente ser cargados en la tabla de hechos.		
Entrada: dpartemitranstransporte		
Criterio de aceptación: Deben ser insertados en la tabla de hechos los siguientes valores dim_vehiculo_id, dim_temporal_dia_id, cantidad_disponibles, cantidad_en_taller, cdt_sem_actual, cdt_sem_proxima, cdt_sem_anterior.		
Resultado: se almacenan los datos en la tabla de hechos hech_balance_transporte.		

4.5 Aporte social y económico

Para los directivos el factor tiempo es de vital importancia para su gestión, hoy en día, la demora puede resultar además de incumplimiento en lo planificado, pérdida de recursos, tanto financieros como humanos.

El desarrollo del Mercado de Datos para la Dirección General Infraestructura y Servicios brinda a los directivos resúmenes de información para conformar reportes comparativos, utilizando los indicadores más comunes de esta Dirección, para apoyarse en la toma de decisiones. Además es una solución que reduce tiempos de espera por parte de los usuarios ofreciendo un incremento en la rapidez de las consultas.

El Mercado de Datos es capaz de ofrecer un acceso integrado, consistente, fiable y rápido a los datos, que permite tomar decisiones basadas en una mejor

información. El principal aporte lo constituye lograr mayor productividad por decisiones correctas en un tiempo más corto lo que también conllevaría un beneficio económico para la provincia.

La utilización del Mercado de Datos implica un impacto positivo sobre los procesos que se llevan a cabo en esta Dirección, dotándola de información estratégica para su mejor desempeño.

Conclusiones del capítulo

En este capítulo se expuso la validación funcional del Mercado de Datos. Arrojándose los siguientes resultados:

1. Se elaboró el Plan de pruebas para probar las funcionalidades del Mercado de Datos Infraestructura y Servicios.
2. Se diseñaron los Casos de Prueba basados en cada uno de los Casos de Uso Informativos.
3. Se aplicaron los Casos de Prueba diseñados para verificar que el Mercado de Datos tuviera la calidad requerida.
4. Se realizaron las pruebas al subsistema de integración arrojando resultados satisfactorios.

CONCLUSIONES

La realización del presente trabajo de diploma propició el estudio acerca del estado actual de los Almacenes y Mercados de Datos a nivel mundial y nacional, y de la importancia que tiene la utilización de los mismos para ayudar en el proceso de la toma de decisiones.

Con la elaboración del trabajo se cumplió el objetivo trazado, que permitió a través de las tareas definidas guiar el proceso de desarrollo del sistema que contribuirá al proceso de la toma de decisiones en la Dirección General Infraestructura y Servicios de la Administración Provincial de Artemisa:

1. Se realizó la Fundamentación Teórica de los Almacenes de Datos.
2. Se caracterizó el proceso de análisis de la información en la Dirección General Infraestructura y Servicios.
3. Se desarrolló el Mercado de Datos Infraestructura y Servicios.
4. Se validó funcionalmente el Mercado de Datos Infraestructura y Servicios.

Recomendaciones

- Poblar el Mercado de Datos con toda la información histórica que posee la Dirección General Infraestructura y Servicios.
- Tener en cuenta las nuevas necesidades de información que presenten los usuarios.
- Desplegar el sistema en la Dirección General Infraestructura y Servicios.

REFERENCIAS BIBLIOGRÁFICAS

Álvarez, Jaime Velázquez. 2010. *Implementación del modelo de datos ArcHydro en la región hidrológica Balsas.* Mexico : s.n, 2010.

Armstrong, Smith. 2006. *Oracle Discoverer 10g Handbook.* San Francisco, California : s.n, 2006.

Business Intelligence+Informática estratégica. 2007. Sinnexus. [En línea] 2007. [Citado el: 2011 de 10 de 2.] http://www.sinnexus.com.

—. 2002. Sinnexus. [En línea] 2002. http://www.sinnexus.com/business_intelligence/olap_avanzado.aspx.

Cañete, Patricio. 2006. Business. Weblog. [En línea] 2 de Agosto de 2006. [Citado el: 15 de 11 de 2011.] Córdoba Argentina. http://www.pcanete.com.ar/leer.asp?idx=414.

Company Headquarters. 2009. Visual Paradigm. *Visual Paradigm.* [En línea] 2009. http://www.visual-paradigm.com/.

Espinosa, Itziar Angoitta. 2007. Data Warehouse para la Gestión de Lista de Eespera Sanitaria. 2007.

ETL-Tools.Info. 2006. Business Intelligence - Almacenes de Datos - ETL. [En línea] 2006. [Citado el: 21 de 10 de 2011.] http://etl-tools.info/es/bi/proceso_etl.htm.

Gerolami, Nicolás. 2011. *Implantación de Data Warehouse Open Free.* Montevideo - Uruguay : s.n., 2011.

González, Yaneisy Pedraza. 2011. Sistema de información de gobierno Mercado de datos Inmigración y extranjería. Ciudad de la Habana : Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas., 2011.

Kimball, Ralph. 2008. *The Data Warehouse Lifecycle Toolkit.* EUA : Wiley Publishing Inc, 2008.

Lanzillotta, Analía. 2012. Mastermagazine. Definición de OLAP – Tecnología OLAP. [En línea] 2007, 2012. http://www.mastermagazine.info/termino/6841.php.

Mendez, Ana Laura Alba. 2009. Arquitectura, Diseño, Construccion, Mantenimiento y Consulta de un Almacen de Datos. 2009.

Nader, Javier. 2003. Sistema de Apoyo Gerencial Universitario. s.l. : Tesis de Magister en Ingeniería del software , 2003.

NEXTEL Engineering. 2010. La Inteligencia de Negocios. [En línea] 2010. [Citado el: 2 de 10 de 2011.] http://www.nexteleng.es.

Ponniah, Paulraj. 2001. *Data Warehousing Fundamentals.* EUA : Wiley Publishing Inc, 2001.

Rivadera, Gustavo R. 2010. *La metodología de Kimball para el diseño de almacenes de.* 2010.

Salazar, Ricardo Luján. 2008. *Data Warehouse para la prestación del servicio público de información estadística.* Mexico : s.n., 2008.

Sánchez, Leopoldo Zenaido Zepeda. 2008. *Metodología para el Diseño Conceptual de Almacenes de Datos.* Valencia : s.n., 2008.

Sanz, Miguel Rodríguez. 2010. Analisis y Diseño de un Data Mart para el Seguimiento Academico de Alumnos en un Entorno Univrsitarioi. 2010.

Sierra, Julio Ernesto Ortiz. 2009. *Diseño e Implementación de un Mercado de Datos para la Oficina Nacional de Estadísticas.* Ciudad de la Habana : Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas., 2009.

Soto, David. 2008. Integración y Calidad de Datos. [En línea] 17 de 7 de 2008. [Citado el: 25 de 2 de 2012.] http://integracionycalidad.blogspot.com/2008/07/migraciones-fusiones-y-adquisiciones.html.

Soto, Esperanza Gil. 2001. *Datawareouse: antecedentes, situacion actual y tendencias.* Santa Cruz de Tenerife : s.n., 2001.

BIBLIOGRAFÍA

Álvarez, Jaime Velázquez. 2010. Implementación del modelo de datos ArcHydro en la región hidrológica Balsas. Mexico : s.n., 2010.

Antunez, Ivette Marrero. 2008. La inteligencia de negocios desde la perspectiva cubana: retos y tendencias. La Habana, Cuba: s.n., 2008.

Armstrong, Smith. 2006. Oracle Discoverer 10g Handbook. San Francisco, California : s.n., 2006.

Business Intelligence+Informática estratégica. 2007. Sinnexus. [En línea] 2007. [Citado el: 2011 de 10 de 2.] http://www.sinnexus.com..

—. 2002. Sinnexus. [En línea] 2002. http://www.sinnexus.com/business_intelligence/olap_avanzado.aspx.

Cañete, Patricio. 2006. Business. Weblog. [En línea] 2 de Agosto de 2006. [Citado el: 15 de 11 de 2011.] Córdoba Argentina. http://www.pcanete.com.ar/leer.asp?idx=414.

Company Headquarters. 2009. Visual Paradigm. Visual Paradigm. [En línea] 2009. http://www.visual-paradigm.com/.

Espinosa, Itziar Angoitta. 2007. Data Warehouse para la Gestión de Lista de Eespera Sanitaria. 2007.

ETL-Tools.Info. 2006. Business Intelligence - Almacenes de Datos - ETL. [En línea] 2006. [Citado el: 21 de 10 de 2011.] http://etl-tools.info/es/bi/proceso_etl.htm.

García, D. Andrés Boza. 2004. Data Warehouse para la gestión por procesos en el sistema. Cancun, Mexico : s.n., 2004.

Gerolami, Nicolás. 2011. Implantación de Data Warehouse Open Free. Montevideo - Uruguay : s.n., 2011.

González, Yaneisy Pedraza. 2011. Sistema de información de gobierno Mercado de datos Inmigración y extranjería. Ciudad de la Habana : Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas., 2011.

Hernández, Ing.Yanisbel González. 2011. Propuesta de Metodología para el desarrollo de Almacenes de Datos en DATEC. Cuidad de la Habana : UCI, Centro de Tecnologías de Gestión de Datos (DATEC), 2011.

Kimball, Ralph. 2008. The Data Warehouse Lifecycle Toolkit. EUA : Wiley Publishing Inc, 2008.

Lanzillotta, Analía. 2012. Mastermagazine. Definición de OLAP – Tecnología OLAP. [En línea] 2007, 2012. http://www.mastermagazine.info/termino/6841.php.

Mendez, Ana Laura Alba. 2009. Arquitectura, Diseño, Construccion, Mantenimiento y Consulta de un Almacen de Datos. 2009.

Nader, Javier. 2003. Sistema de Apoyo Gerencial Universitario. s.l. : Tesis de Magister en Ingeniería del software , 2003.

NEXTEL Engineering. 2010. La Inteligencia de Negocios. [En línea] 2010. [Citado el: 2 de 10 de 2011.] http://www.nexteleng.es.

Ponniah, Paulraj. 2001. Data Warehousing Fundamentals. EUA : Wiley Publishing Inc, 2001.

Rivadera, Gustavo R. 2010. La metodología de Kimball para el diseño de almacenes de. 2010.

Salazar, Ricardo Luján. 2008. Data Warehouse para la prestación del servicio público de información estadística. Mexico : s.n., 2008.

Sánchez, Leopoldo Zenaido Zepeda. 2008. Metodología para el Diseño Conceptual de Almacenes de datos. Valencia : s.n., 2008.

Sanz, Miguel Rodríguez. 2010. Analisis y Diseño de un Data Mart para el Seguimiento Academico de Alumnos en un Entorno Univrsitarioi. 2010.

Sierra, Julio Ernesto Ortiz. 2009. Diseño e Implementación de un Mercado de Datos para la Oficina Nacional de Estadísticas. . Ciudad de la Habana : Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas., 2009.

Soto, David. 2008. Integración y Calidad de Datos. [En línea] 17 de 7 de 2008. [Citado el: 25 de 2 de 2012.] http://integracionycalidad.blogspot.com/2008/07/migraciones-fusiones-y-adquisiciones.html.

Soto, Esperanza Gil. 2001. Datawareouse: antecedentes, situacion actual y tendencias. Santa Cruz de Tenerife : s.n., 2001.

Thomsen, Erik. 2002. OLAP Solutions. s.l. : John Wiley & Sons, Inc., 2002.

Wolff, Carmen Gloria. 2002. Modelamiento Multidimensional. [En línea] 28 de 8 de 2002. [Citado el: 19 de 11 de 2011.] http://www.inf.udec.cl/revista/edicion4/cwolff.htm.

ANEXOS

Anexo 1: Entrevista realizada a los trabajadores de la Dirección General Infraestructura y Servicios:

1. ¿Cuáles son los objetivos de su organización? ¿Que se está tratando de resolver? ¿Cuáles son las prioridades de la entidad que deben resolverse?
2. ¿Existen categorías de la información dentro del negocio? ¿Es posible dividir las actividades que se realizan en el negocio por sector, rama, departamento, dirección según el tipo de información que se maneja?
3. ¿Cuál es la información o los tipos de reportes que actualmente se puede obtener en el negocio? ¿Cuáles son las claves del negocio en riesgo actuales?
4. ¿Con qué frecuencia se obtiene información?
5. ¿Qué cantidad de información en general se maneja?

6. ¿Cuáles son las métricas para medir el éxito? ¿Cómo obtienes los indicadores (métricas) para monitorear los procesos?
7. ¿Qué criterios se tienen en cuenta para analizar los indicadores?
8. ¿Qué análisis le gustaría realizar sobre esos indicadores? ¿Existen posibilidades de mejora a su método/ proceso actual?
9. ¿Existen eventos externos que es interesante analizar sobre esos indicadores (actividades en la universidad, días feriados, etc.)?
10. ¿Existen cortes de información o peticiones más comunes? ¿Qué tipo de análisis a la medida suele realizar con más frecuencia?
11. ¿Cómo le gustaría que se le presentaran los datos en la pantalla? (gráficos (tipos), tablas, indicadores en la parte superior, etc.)
12. Describa sus fuentes de datos (u otra clase de dimensión).

GLOSARIO DE TÉRMINOS

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

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

BI: Inteligencia del Negocio (del inglés, Business Intelligence). Conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa.

Consistencia de los datos: significa que los datos deben usar una convención estándar que determine su significado.

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

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

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

EPL: Licencia Pública Eclipse del inglés Eclipse Public License. Es una licencia de software de código abierto utilizada por la Fundación Eclipse para su software.

GPL: Licencia Pública General (del inglés, General Public License). Es la licencia de software libre más utilizada. La GPL otorga a los beneficiarios de un programa de ordenador de los derechos de la definición de software libre.

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

Jpivot: conjunto de bibliotecas configurables JSP que permiten mostrar tablas y gráficos, para mostrar la navegación típica de los entornos OLAP.

JSP: del inglés JavaServer Pages, es una tecnología Java que permite generar contenido dinámico para web, en forma de documentos HTML, XML o de otro tipo.

MDX: Expresiones multidimensionales (MDX, Multidimensional Expressions) es un lenguaje de consulta para bases de datos OLAP.

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

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

ORDBMS: Es el acrónimo en inglés de Objeto de Sistema de Gestión de Base de Datos (Object Relational Data Base Management System). Es un Sistema de

Gestión de Base de datos similar a una base de datos relacional, pero con un modelo de base de datos orientada a objetos.

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

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

Software Libre: Software que, una vez obtenido, puede ser usado, copiado, estudiado, modificado y redistribuido libremente. El software libre suele estar disponible gratuitamente en Internet, o a precio del costo de la distribución a través de otros medios.

TCP/IP: Protocolo que permite establecer una conexión e intercambiar datos. Garantiza la entrega de datos y que los paquetes sean entregados en el mismo orden que fueron enviados.

Transacciones: Una transacción es una unidad de la ejecución de un programa que accede y, posiblemente, actualiza varios elementos de datos. Una Transacción está delimitada por instrucciones de inicio y fin (la transacción consiste en todas las operaciones que se ejecutan entre inicio y fin de la misma).

XML: Lenguaje de Marcas Extensible (del inglés, eXtensible Markup Language). Es un metalenguaje extensible de etiquetas. Permite la compatibilidad entre sistemas para compartir la información de una manera segura, fiable y fácil.