

Sistema para apoyar la toma de decisiones en la Dirección General Infraestructura y Servicios.

Sonia Fernández Henríquez

Universidad de las Ciencias Informáticas, La Habana, Cuba, shenriquez@uci.cu

Diana Rosa Prieto del Rio

Instituto Superior Politécnico José Antonio Echeverría, La Habana, Cuba, drprieto@cemat.cujae.edu.cu

Marta Rodríguez Freire

Universidad de las Ciencias Informáticas, La Habana, Cuba, mfreire@uci.cu

RESUMEN

En la actualidad 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 y de esta forma apoyar la toma de decisiones.

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.

En el presente trabajo se aborda sobre el desarrollo de un Mercado de Datos para la Dirección General Infraestructura y Servicios de la Administración Provincial de Artemisa que contribuya al proceso de la toma de decisiones en dicha dirección, como factor de éxito para la provincia de Artemisa. Se describe de forma general el sistema y la implementación del mismo.

Palabras claves: Almacenes de Datos, Mercado de Datos, Toma de decisiones.

ABSTRACT

In the country there is still ignorance on the subject of Business Intelligence and Data Warehouses, because several organizations are not fully aware of the capabilities that these technologies can provide. Despite this over time institutions are feeling increasingly attracted by the benefits associated with Business Intelligence solutions.

This work is about the development of Data Mart for Infrastructure and Services General Direction of the Provincial Administration of Artemisa that contributes to the decision making process in this direction as a success factor for the province of Artemisa. It describes the system and the implementation thereof.

Keywords: Data Warehouses, Data Mart, Decision making.

1. INTRODUCCIÓN

Actualmente en las empresas se necesita tomar decisiones en respuesta a las exigencias de las actividades y funciones que sus trabajadores realizan. 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 con los cuales 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.

Actualmente la Dirección General Infraestructura y Servicios de la Administración Provincial de Artemisa presenta dificultades para el análisis de los indicadores a medir en los procesos que se llevan a cabo en esta dirección, 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.

La presente investigación tiene como objetivo desarrollar un Mercado de Datos que contribuya a la centralización, consistencia y disponibilidad de la información en la Dirección General Infraestructura y Servicios.

2. MARCO TEÓRICO

Los Almacenes de Datos según Ralph Kimball, "...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.

Los Mercados de Datos son Almacenes de Datos orientados a temas específicos o aplicaciones específicas y contienen datos de sólo una línea del negocio como puede ser ventas o marketing. (Hobbs et al., 2005)

Los cubos multidimensionales son estructuras que representan los datos como una matriz en la cual sus ejes corresponden a los criterios de análisis y en los cruces se encuentran los valores a analizar. (Gerolami et al., 2011)

Elementos asociados a los cubos multidimensionales:

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

Dimensiones: son perspectivas o entidades respecto a las cuales una organización quiere mantener sus datos organizados. (Mendez, 2009)

Hechos: Los hechos son criterios de análisis que contienen las medidas numéricas que serán utilizados por los analistas del negocio para apoyar el proceso de toma de decisiones. Contienen datos cuantitativos necesarios para el análisis. (Gerolami et al., 2011)

Jerarquías: están representadas por un ordenamiento lógico dentro de la dimensión, se encuentran formadas por los diferentes tipos de relaciones entre los atributos de una misma dimensión. (Marroquín et al., 2007)

2.1 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.

El proceso de Extraer, Transformar y Cargar (ETL) 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)

2.2 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.

3. HERRAMIENTAS Y METODOLOGÍAS SELECCIONADAS PARA EL DESARROLLO DE LA SOLUCIÓN.

En la presente investigación se decidió utilizar en su mayoría herramientas de código libre, siguiendo la política de software libre del país y la Universidad de las Ciencias Informáticas.

Como herramienta de modelado se seleccionó el Visual Paradigm 6.4 puesto que es una herramienta que soporta el ciclo de vida completo del desarrollo de software, es multiplataforma además permite realizar Ingeniería Inversa de bases de datos desde Sistemas Gestores de Bases de Datos existentes a diagramas de Entidad-Relación.

Para realizar el proceso de ETL y presentar los datos para el análisis al usuario se eligieron las herramientas de la Suite de Pentaho. Estas herramientas son todas libres, multiplataforma y fáciles de usar.

Como Sistema Gestor de Base de Datos (SGBD) se optó por la herramienta PostgreSQL 9.1 la cual es uno de los SGBD más usados en la actualidad y como interfaz gráfica de este SGBD se utilizó PgAdmin III 1.14, siendo la más completa y popular con licencia libre. Soporta todas las características de PostgreSQL y facilita enormemente su administración.

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 de las Ciencias Informáticas, 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.

4. DESCRIPCIÓN DE LA SOLUCIÓN DESARROLLADA

La solución desarrollada está diseñada de acuerdo a las necesidades de análisis de los indicadores que se manejan en la Dirección General Infraestructura y Servicios.

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. El sistema cuenta con 3 subsistemas como se muestra en la figura 1: Subsistema de integración, Subsistema de almacenamiento, Subsistema de visualización.

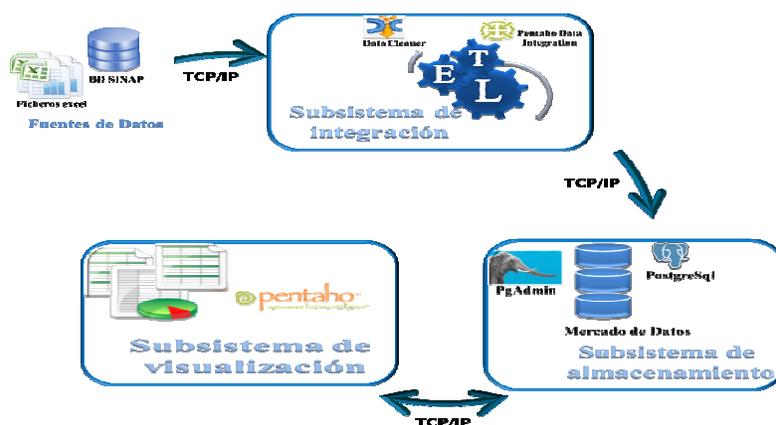


Figura 1: Representación de la Arquitectura.

En el subsistema de integración es donde se efectúan todas las tareas desde que se extraen los datos de las fuentes hasta que se puebla el Mercado de Datos. Agrupa las tareas relacionadas con la extracción, integración, limpieza y carga de los datos al Mercado de Datos.

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.

El subsistema de visualización es la capa con la cual interactúa el usuario final. Tiene como finalidad mostrar los datos almacenados de forma útil a través de las distintas herramientas de Pentaho. A través de consultas a los cubos de datos, se obtiene la información requerida para ser presentada para la visualización de los reportes y las vistas de análisis.

Para responder a las necesidades de información de los usuarios finales se identificaron 24 tablas de hechos, 20 tablas de dimensiones y 148 medidas a partir de lo cual se conformó el modelo de datos de la solución. Para el modelado de los datos se siguió el modelo constelación de hechos, debido a que contribuye a la reutilización de las dimensiones ya que una misma dimensión puede ser utilizada por varias tablas de hechos.

Para lograr que la información esté almacenada, preparada e integrada en el Mercado de Datos, fue preciso realizar la implementación del proceso de ETL. Se implementaron 2 tipos de cargas: cargas históricas a partir de la información en ficheros Excel con la cual ya contaba la Dirección General Infraestructura y Servicios y cargas incrementales teniendo como fuente de datos la base de datos del Sistema Informativo de la Administración Provincial. Para realizar este proceso se siguió el flujo de actividades que se muestra en la figura 2. Inicialmente

se extraen 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 (en el caso de las cargas incrementales). 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 el caso de las cargas históricas los errores se almacenaron en ficheros Excel), 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 (sólo en el caso de las cargas incrementales).

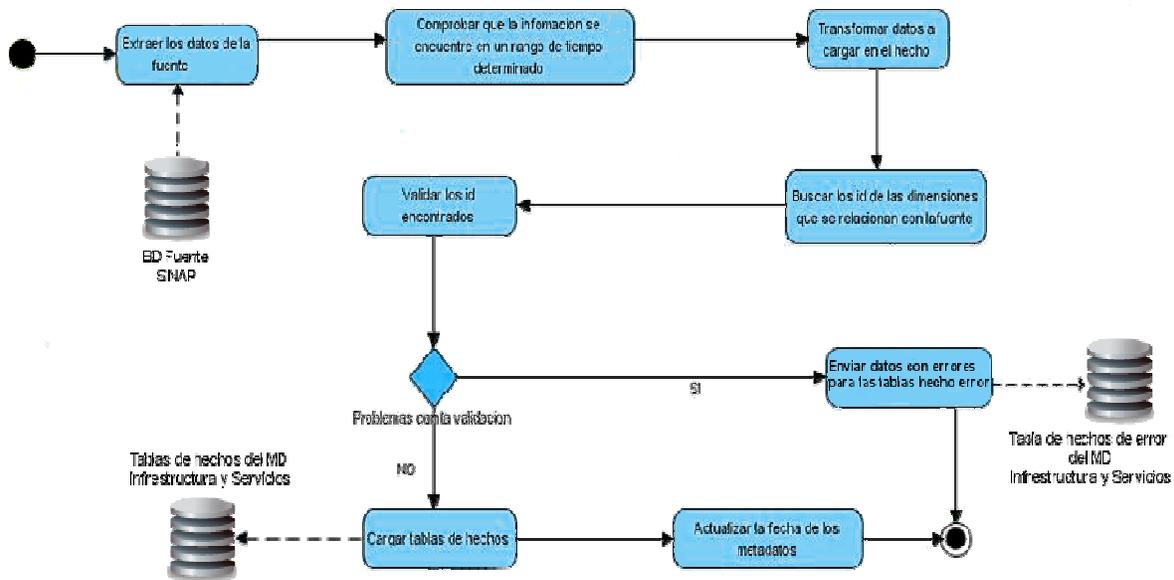


Figura 2: Proceso ETL para las cargas incrementales.

La solución cuenta con una transformación para cada tabla de hechos. A continuación se muestra una de las transformaciones realizadas para cargar una de las tablas de hechos del Mercado a partir de la base de datos del SINAP. Primeramente se extraen los datos de la tabla que guarda la información correspondiente a las dietas. Luego se realizan una serie de transformaciones a los datos. Luego se buscan los identificadores de las tablas de dimensiones *dim_dietas*, *dim_temporal_mes* y *dim_municipio*. Finalmente los datos se filtran por los identificadores mencionados anteriormente. Si no existe ningún problema los datos son cargados satisfactoriamente en la tabla de hechos *hech_balance_dietas*. En caso contrario los datos son almacenados en la table de hecho error asociada a esta tabla de hechos.

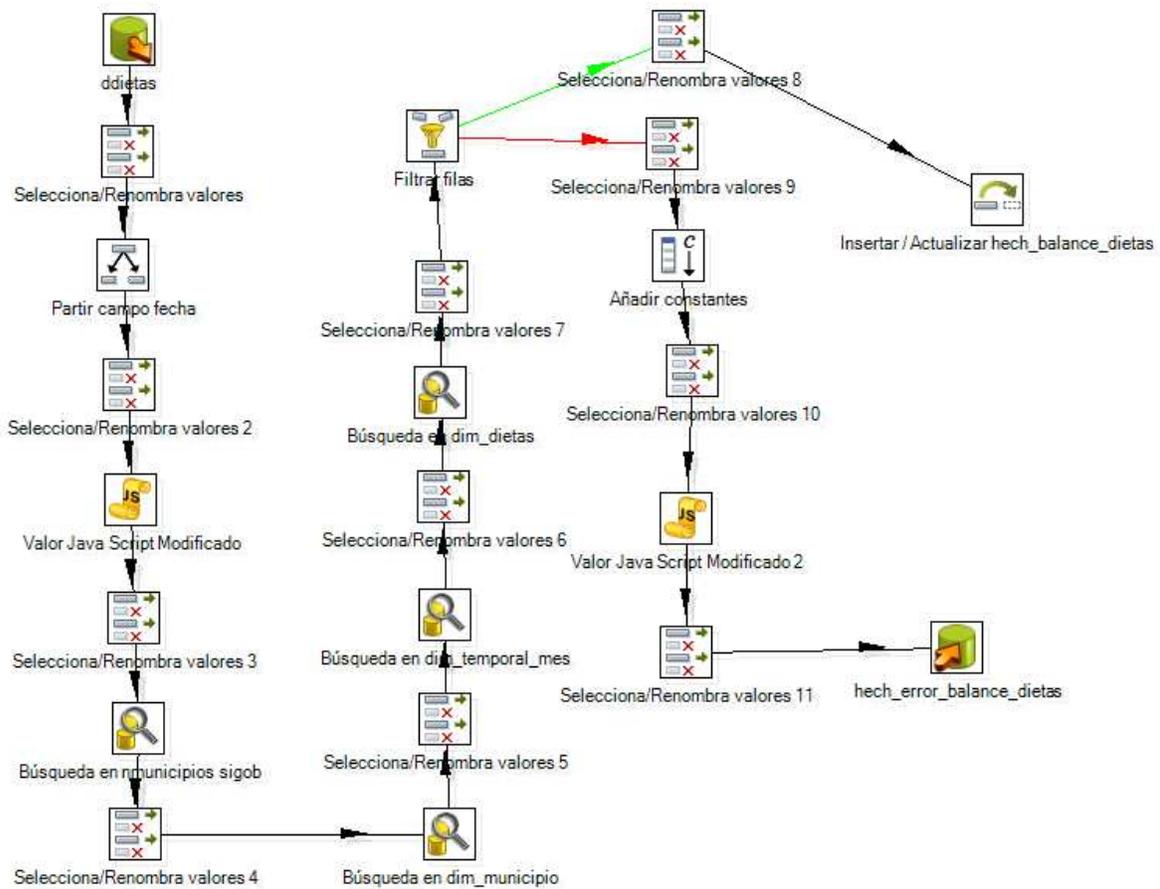


Figura 3: Transformación para cargar los datos a la tabla de hechos balance de dietas.

Una vez realizadas todas las transformaciones necesarias se implementaron los Trabajos con el objetivo de organizar el proceso de carga al Mercado de Datos. La solución cuenta con un trabajo por cada tabla de hechos. Seguidamente se describe de forma general este proceso como se muestra en la figura 3: se obtienen los datos asociados a la tabla fuente de donde se van a cargar los mismos. Se obtiene la fecha de la computadora donde se encuentra la tabla fuente y la fecha de la última vez que los datos fueron cargados al Mercado de Datos. Se ejecuta la transformación correspondiente a la tabla de hechos que se vaya a cargar y por último se actualiza la tabla de metadatos.

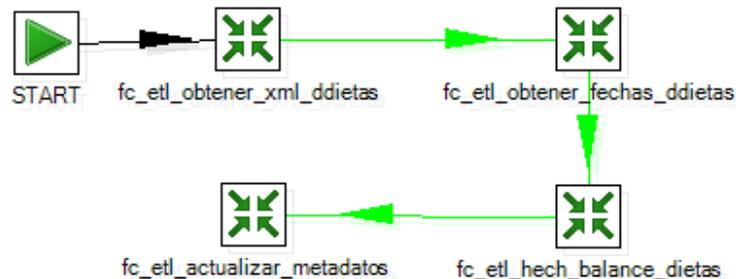


Figura 4: Trabajo realizado para cargar para cargar los datos a la tabla de hechos balance de dietas.

Además la solución cuenta con un trabajo general que ejecuta a los demás en dependencia de la frecuencia de carga (diario, semanal, mensual) de la información que se maneja en la Dirección General.

Para lograr la correcta presentación de los datos al usuario final se hace necesario crear los cubos multidimensionales. En el sistema se definieron 25 cubos de datos.

Para la creación de los cubos es preciso definir la tabla de hechos, las dimensiones que se relacionan con dichas tablas de hechos, las medidas y los miembros calculables. Estos no son más que los atributos que se calculan a partir de las medidas ya definidas. Para poder realizar el análisis de los cubos, estos deben publicarse dentro de la plataforma Pentaho. A continuación se presenta la estructura general de uno de los cubos utilizados dentro de la solución propuesta.



Figura 5: Cubo de datos.

Una vez que los cubos han sido publicados en la herramienta Pentaho BI-Server es posible mostrarle al usuario el resultado del análisis realizado, a través de los reportes y las vistas de análisis previamente definidas. El Mercado de Datos Infraestructura y Servicios cuenta con 48 vistas de análisis y 23 reportes útiles con el objetivo de satisfacer las necesidades de información de la Dirección General Infraestructura y Servicios. Además de los reportes definidos 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.

A continuación se describen dos de los requisitos de información que integran la solución:

Obtener de la situación de los embalses el tiempo de bombeo, el caudal de bombeo, el volumen bombeado, por estación de bombeo y por período.

		Medidas		
Periodo	Estación de bombeo	● Caudal de bombeo (m3/s)	● Volumen bombeado	● Tiempo de bombeo (horas)
<input type="checkbox"/> Ver todos	<input type="checkbox"/> Ver todos	42,82	566495,00	867,00
<input type="checkbox"/> 2011	<input type="checkbox"/> Ver todos	42,82	566495,00	867,00
	Pinillos	42,82	566495,00	867,00
<input type="checkbox"/> 1er Semestre	<input type="checkbox"/> Ver todos	34,00	34,00	45,00
	Pinillos	34,00	34,00	45,00
<input type="checkbox"/> 2do Semestre	<input type="checkbox"/> Ver todos	8,82	566461,00	822,00
	Pinillos	8,82	566461,00	822,00

Figura 6: Vista de análisis del Mercado de Datos.

Obtener de la situación de los embalses la capacidad total de embalse de cada complejo, el por ciento de embalsado de cada complejo hidráulico, el volumen embalsado, por complejo y por período.

Complejo	Capa. Embalse	Vol Embalsado	%
complejo oeste	156	28	17,9
complejo este	1.189	127	10,7

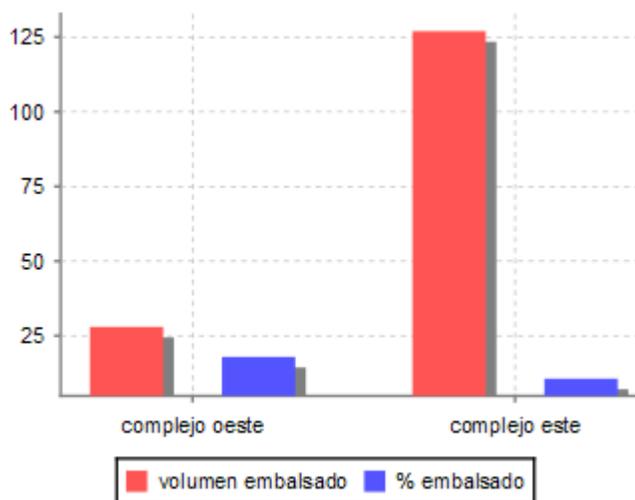


Figura 7: Reporte del Mercado de Datos.

5. RESULTADOS OBTENIDOS

El desarrollo del Mercado de Datos para la Dirección General Infraestructura y Servicios ofrece a los directivos resúmenes de información para conformar reportes comparativos, utilizando los indicadores más frecuentes 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.

La solución tiene entre sus funcionalidades las siguientes: autenticar usuario, mostrar el gráfico asociado al reporte seleccionado, configurar las diferentes vistas de análisis por las que se muestra un reporte, realizar la extracción, transformación y carga de los datos, además le permite al usuario elaborar sus propios reportes.

6. CONCLUSIONES

El desarrollo del presente trabajo permitió realizar un estudio acerca de los fundamentos teóricos de los Almacenes de Datos. Se cumplió con el objetivo del trabajo pues se desarrolló un Mercado de Datos para la Dirección General Infraestructura y Servicios, el cual 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.

7. REFERENCIAS

- Armstrong, Smith. 2006. Oracle Discoverer 10g Handbook. San Francisco, California: s.n, 2006.
- Cañete, Patricio. 2006. Business. Weblog. [en línea] 2 de Agosto de 2006. Disponible en Web: <http://www.pcanete.com.ar/leer.asp?id=41> [consultado noviembre de 2011]

- ETL-Tools.Info. 2006. Business Intelligence - Almacenes de Datos - ETL. [en línea] 2006. Disponible en Web: http://etl-tools.info/es/bi/proceso_etl.htm. [ref. octubre de 2011]
- Hobbs, Lilian y otros. 2005. Oracle Data Base 10g Data Warehousing. EUA: ELSEVIER Digital Press, 2005.
- Gerolami, Nicolás; Revello Esteban; Venzail German. “Implantación de Data Warehouse Open Free”. Tutores: Nelson Calero, Lorena Etcheverry, Ariel Sabiguero. Proyecto de grado, Facultad de Ingeniería, Instituto de Computación, Montevideo, Uruguay, 2011.
- Reyes Marroquín, Mario Roberto; Rosales Tejeda, Pablo Augusto. “Desarrollo de un Datamart de Información académica de estudiantes de la Escuela de Ciencias y Sistemas de la Facultad de Ingeniería de la USAC”. Director: Ing. Jorge Armín Masariegos. Tesis para optar por el título de Ingeniero en Ciencias y Sistemas, Universidad de San Carlos de Guatemala, Guatemala, 2007.
- Rivadera, Gustavo R. “La metodología de Kimball para el diseño de almacenes de datos (Data warehouses)”. [en línea] ebookbrowse [ref. junio 2011]. Disponible en Web: <http://ebookbrowse.com/5-p56-rivadera-formateado-pdf-d139304020> [consultado febrero 2012].
- Mendez Alba, Ana Laura. “Arquitectura, Diseño, Construcción, Mantenimiento y Consulta de un Almacén de Datos”. Director: M.C.C Ericka Janet Rechy Ramírez. Tesis para optar por el título de Licenciado en Sistemas Computacionales Administrativos, Universidad Veracruzana, Xalapa, 2009.

AUTORIZACIÓN Y RENUNCIA

Los autores autorizan a LACCEI para publicar el artículo en las actas de congresos. Ni los editores ni LACCEI son responsables ni por el contenido ni por las implicaciones de lo que se expresa en el documento.