Ediciones pensando el futuro
Editions thinking the future
Informática 2013
XV CONVENCIÓN Y FERIA INTERNACIONAL
15TH INTERNATIONAL CONVENTION AND FAIR
www.informaticahabana.com
www.informaticahabana.cu
ISBN: 978-959-7213-02-4

SLD080 MERCADO DE DATOS PARA UNA DIRECCIÓN DE SALUD EN CUBA

SLD080 DATA MART FOR HEALTH MANAGEMENT IN CUBA

Geidy Acosta Méndez[1], Disnayle Jorge Chacón[2]

1 Universidad de las Ciencias Informáticas (UCI), Cuba, geidyam@uci.cu, Ave 95 #3215 %32 y 34 Cotorro La Habana 14000

2 Universidad de las Ciencias Informáticas (UCI), Cuba, djorge@uci.cu

RESUMEN: *El presente trabajo se enmarca en el tema de los almacenes de datos, los mercados de datos y su utilización para los análisis estadísticos de la información en una dirección de salud. La metodología utilizada es Proceso de Desarrollo en la Línea Soluciones de Almacenes de Datos Inteligencia de Negocio la cual se basa en las ventajas de las metodologías de Kimball y Inmon, en el desarrollo del trabajo se detallan topología, herramientas, especificación de requerimientos, para lograr un buen diseño e implementación de los procesos de integración y análisis de datos de la solución. Como resultado se obtiene un mercado de datos poblado disponible para realizar análisis detallados, y se tiene la estructura del modelo dimensional que comprende: las dimensiones, las jerarquías, las tablas de hechos y las medidas necesarias para proceder con los cálculos y análisis estadísticos. De igual manera, la solución incluye las políticas de seguridad, así como las pruebas para la validación del mercado.*

Palabras Clave: Almacén de datos, dimensión, mercado de datos.

Abstract: *This work is framed in the theme of data warehouses, data marts and their use for statistical analysis of information in an direction health. The methodology used in the Development Process Line Data Warehousing Solutions Business Intelligence which is based on the advantages of Kimball and Inmon methodologies, in the development of work are topology, tools, requirements specification detailed, to achieve good design and implementation of the integration processes and analysis of the solution. The result is a market populated data available for detailed analysis and has dimensional model structure comprising: dimensions, hierarchies, fact tables and the necessary steps to proceed with the calculations and statistical analysis. Similarly, the solution includes security policies and testing for market validation.*

KeyWords: Data mart, data warehouse, dimension.

1. INTRODUCCIÓN

La gestión de la información se convierte en una forma de marcar la diferencia y hacer ventaja competitiva en un mundo globalizado. En este sentido, simples formatos y registros son calificados como herramientas básicas de recopilación de información en especial necesidades de clientes, quejas, reclamos e incluso nuevos servicios solicitados. Esto ayuda también a la incorporación de factores de innovación en las empresas, como nuevas tecnologías de la información y de la comunicación, las cuales hacen más sencillo la incorporación de Bases de Datos.

Con el transcurrir del tiempo estas Bases de Datos comenzaban a generar gran acumulación de información. Los directivos de tales empresas y negocios se dieron cuenta que estas podían tener un fin útil al estar reflejadas la mayoría de sus operaciones comerciales durante los ciclos de negocios propios del mercado.

Por tanto pensaron en lo ideal que sería unificar las diferentes fuentes de información de las cuales disponían en un único lugar, al que solamente se le incorporaría información relevante sobre la base de una estructura organizada, integrada, lógica, dinámica y de fácil explotación. La respuesta a esto fueron los Almacenes de Datos.

Los Almacenes de Datos son una tecnología que ha tenido un gran impacto en el medio informático y empresarial desde el momento en que surgió, antes de su nacimiento solamente se trabajaban con datos "fríos", proporcionados por las Bases de Datos y si se quería hacer un análisis empresarial para el mejoramiento de esta empresa, habían ciertos datos que simplemente eran imposibles de obtener y por lo que el análisis era casi improductivo.

En Cuba existen sistemas informáticos especializados en el control y gestión de almacenes, pero no se adaptan a las particularidades de todas las empresas o estas se ven imposibilitadas a adquirirlos por no contar con presupuesto para ello. En este contexto se pretende desarrollar un sistema informático, encaminado a satisfacer las necesidades de gestión de la información existentes en una empresa, una institución o el país en general.

A continuación se describen los problemas que existen en una dirección de salud: No existe un estándar específico para el almacenamiento de los datos, existen inconsistencias en los datos que se manejan, ya que no están organizados de la forma más óptima, el análisis de la información resulta engorroso cuando se trata de grandes volúmenes, la seguridad de los datos se encuentra comprometida, no están definidos niveles de accesibilidad en la información que se maneja, los datos no están integrados, ya que existen referencias a la misma información usando diferente codificación.

Esto dificulta a la entidad el análisis y accesibilidad de los volúmenes de información generados durante sus procesos de trabajo con la agilidad y rapidez necesaria, poniendo en riesgo la seguridad de los modelos informativos que se manejan y de esta manera ralentizando el proceso de la toma de decisiones.

En función de dar solución a las problemáticas planteadas se traza como Objetivo General: Desarrollar un Mercado de Datos para mejorar la toma de decisiones en una Dirección de Salud.

2. ALMACENES DE DATOS

Un Almacén de Datos (ADs) es un repositorio de datos de fácil acceso, alimentado de numerosas fuentes, transformadas en grupos de información sobre temas específicos de negocios, para permitir nuevas consultas y análisis. Es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes, para luego procesarla permitiendo su análisis desde diferentes perspectivas y con grandes velocidades de respuesta. La creación de un ADs representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de inteligencia de negocios [1].

2.1 Principales ventajas y desventajas

Ventajas

a. Brindan rentabilidad en las inversiones realizadas para su creación.
b. Aumenta la competitividad en el mercado.
c. Aumenta la productividad de los técnicos de dirección.
d. Los ADs hacen más fácil el acceso a una gran variedad de datos a los usuarios finales.

Desventajas

a. Sub-valoración del esfuerzo necesario para su diseño y creación.
b. Sub-valoración de los recursos necesarios para la captura, carga y almacenamiento de datos.
c. Incremento continuo de los requisitos de los usuarios.
d. A lo largo de su vida los ADs pueden suponer altos costos. El AD no suele ser estático. Los costos de mantenimiento son elevados [2].

2.2 Mercado de datos (MDs). Características y Metodología

Los MDs son un subconjunto de datos de un AD donde se almacenan la mayoría de las actividades de análisis que en el entorno de Inteligencia de Negocio se llevará a cabo.

La visión de Inmon se basa en un enfoque descendente, propone construir primero el AD, y a partir de este los MDs. Plantea la creación de un repositorio de datos corporativo como fuente de información consolidada, persistente, histórica y de calidad.

A diferencia de la anterior, la propuesta de Kimball se basa en dividir el mundo de Inteligencia de Negocio entre los hechos y las dimensiones, ésta es eficaz y conduce a una solución completa en un corto período de tiempo. Además, tiene abundante documentación y se puede encontrar una respuesta a casi todas las preguntas que se puedan tener. Entre sus características principales, está el hecho de poseer una arquitectura ascendente, plantea que se debe crear por cada departamento un conjunto de MDs independientes orientados a los temas que estén relacionados con él.

2.2.1 Metodología a utilizar en el Mercado de Datos Salud

La metodología que emplearon los autores para la implementación del Mercado de Datos Salud es la de Proceso de Desarrollo en la Línea Soluciones de Almacenes de Datos Inteligencia de Negocio la que posee las siguientes ventajas:

- La técnica posee una gran cantidad de documentación y generalmente se puede encontrar una respuesta a casi todas las problemáticas que puedan presentar.
- Esta metodología de dividir el mundo de BI entre el hecho y las dimensiones es muy eficaz y conduce a una solución completa en un tiempo razonable.
- Es iterativo, donde se construye una pieza a la vez (MDs) garantizando mayor velocidad de respuesta a los clientes.
- La forma de almacenar la información es de fácil entendimiento por parte del usuario lo que permite mayor comprensión para el análisis de los datos que se encuentran integrados.
- Es una metodología resistente y adaptable ante los cambios.

2.3 Modos de almacenamientos de datos

La tecnología de Procesamiento Analítico en Línea OLAP (Online Analytical Processing) permite un uso más eficaz de los ADs para el análisis de datos en línea, proporciona respuestas rápidas a consultas analíticas complejas e iterativas utilizada generalmente para sistemas de ayuda para la toma de decisiones, presenta los datos a los usuarios a través de un modelo de datos intuitivo y natural.

Existen tres modelos para el proceso analítico en línea (OLAP) de la información: ROLAP, MOLAP y HOLAP. El proceso de análisis se realiza de igual forma lo que varía en uno y otro caso es 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 [3].

Características

	MOLAP	ROLAP	HOLAP
Almacenamiento de las Agregaciones	Modelo Multidimensional	Base de datos relacional	Modelo Multidimensional
Almacenamiento de los datos	Modelo Multidimensional	Base de datos relacional	Base de datos relacional
Facilidad de Creación	Sencillo	Muy Sencillo	Sencillo
Velocidad de respuesta	Buena	Regular o Baja	Buena para consultas que posean agregaciones, Regular para datos de bajo nivel
Escalabilidad	Problemas de escalabilidad	Son más escalables	
Recomendados para	Cubos con uso frecuente	Datos que no son frecuentemente usados	Si el cubo requiere una rápida respuesta

Figura. 1: Comparación de las principales características de los tres modelos para el proceso analítico en línea.

En el presente trabajo se empleará Procesamiento Analítico Relacional en Línea (ROLAP).

2.4 Integración de datos

ETL - este término viene de inglés de las siglas Extract-Transform-Load que significan Extraer, Transformar y Cargar organizando 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 AD, limpiarlos y cargarlos en otra Base de Datos [4].

Etapas del proceso de integración de datos

Debido a que los datos deberán ser extraídos, transformados, limpiados y cargados desde el conjunto de archivos DBF hacia el MD, es imprescindible conocer como se realizarán cada una de estas actividades.

Extracción: Obtención de la información de las distintas fuentes tanto internas como externas.

Transformación: Luego de realizarse el proceso de extracción los datos provenientes de las diferentes fuentes pueden ser incoherentes, tener errores o estar incompletos. Con esto se busca obtener datos 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: Organización y actualización de los dtos 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 AD [5].

2.5 Herramientas utilizadas

Las herramientas para la integración de datos son muy útiles para que el proceso de ETL concluya con los resultados esperados, su uso garantiza: ganancias en términos de tiempo y total fiabilidad de los datos.

Pentaho Data Integration 4.0.1

Es de formato abierto y de fácil lectura para los XML que recogen transformaciones, tareas programadas y un repositorio relacional de metadatos ETL.

- Es aplicable a diversos tipos de Base de Datos (SQL server, PostgreSQL, MySQL, Microsoft Access, etc.).
- Posee facilidad para la importación y exportación de datos de un formato a otro cualquiera.
- Su principal fortaleza es la posibilidad que brinda de ser extensible mediante pluggins [6].

Pentaho Schema Workbench 3.2.0

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 [7].

Pentaho BI server 3.6.0

La aplicación más conocida de la Plataforma de BI es la Pentaho BI Server que funciona como una red basada en sistema de gestión de informe, el servidor de integración de aplicaciones y un motor de flujo de trabajo ligero (secuencias de acción.) Está diseñado para integrarse fácilmente en cualquier proceso de negocio.

Pentaho Report Designer

El Pentaho Report Designer es una herramienta independiente que forma parte de la unidad de reportes de Pentaho (Pentaho Reporting), que simplifica el proceso de generación de reportes, permitiendo a los diseñadores de reportes crear rápidamente informes sofisticados y ricos visualmente basados en el proyecto de reportes de Pentaho JFreeReport. El diseñador de reportes ofrece un entorno gráfico familiar, con herramientas intuitivas y fáciles de utilizar, y una estructura de reporte bastante acertada y flexible para darle libertad al diseñador de generar reportes que se adapten totalmente a su gusto y necesidad.

Servidor Web Apache Tomcat

Tomcat es un servidor web con soporte de Servlets y JSPs. Es usado como servidor web autónomo en entornos con alto nivel de tráfico y alta disponibilidad. Dado que el mismo fue escrito en Java, funciona en cualquier sistema operativo que disponga de la máquina virtual Java. En él los usuarios disponen de libre acceso a su código fuente y a su forma binaria en los términos establecidos en la Apache Software Licence. La primera distribución de Tomcat fue la versión 3.0. La versión más reciente es la 7.0.

2.6 Definición de negocios

La Dirección de Salud tiene el control de toda la información con carácter relevante que se maneja en cada uno de los departamentos que componen el mismo. La información estadística se encuentra almacenadas en los modelos que se trabajan, los cuales se llenan por especialistas de cada departamento y posteriormente revisados por el director de cada departamento para evitar la mayor cantidad de errores posibles.

2.7 Reglas del negocio

Las reglas del negocio se identificaron en el levantamiento de información y en el análisis de las fuentes. Dichas reglas son una entrada fundamental para los procesos de diseño del almacén, ETL y BI.

- Para los campos en blanco, se le introduce el valor 0.
- Los campos predefinidos se comprueban con los nomencladores.
- No debe permitirse cargar información de una fecha anterior cuando se carga una actual.
- No debe permitirse cargar información de la misma fecha una vez que esta haya sido cargada.
- Una vez cargados los datos en el almacén, no pueden existir campos nulos.

2.8 Requisitos Informativos

Los requisitos informativos son la cantidad de reportes internos de cada sistema. En el mercado de datos para una Dirección de Salud se detectaron 53 requisitos los cuales se dividieron por temas de análisis, a continuación como ejemplos se presen-

tan 2 requisitos informativos:

- Obtener del modelo Plan Quirúrgico: el plan total de operaciones, el plan de operaciones mayores, el plan de operaciones menores, el plan de operaciones electivas, el plan de operaciones ambulatorias, el plan de operaciones de ingresados, el plan de operaciones urgentes, el plan de operaciones de mínimo acceso, el plan de operaciones menores electivas, el plan de operaciones menores urgentes por hospitales y por año.
- Obtener del modelo Plan de Operaciones: el total de operaciones realizadas y la diferencia entre estas y el plan de cumplimiento por cada uno de los hospitales y por año.

2.9 Propuesta del sistema

Al concluir el análisis y de acuerdo a las necesidades de los usuarios se propone realizar un sistema que brinde las siguientes funciones:

- Lograr un análisis de la información con el propósito de la proyección de la información para la toma de decisiones.
- Donde se centralicen los datos de los sistemas fuentes, guardarlos durante períodos de tiempo extensos permitiendo el acceso a datos históricos por años lo que proporciona un nivel de análisis basado en experiencias.
- Proporcionar al usuario una interfaz consolidada única para los datos, que hace más fácil el trabajo con las consultas para la toma de decisiones.

2.10 Implementación de la base de datos

El modelado de datos es uno de los elementos más importantes a la hora de iniciar el desarrollo de cualquier proyecto. La verdadera esencia de una aplicación reside en esta estructura. Cuando se diseña el modelo dimensional, el mismo se transforma a un modelo físico, del cual se genera el script de la base de datos, y es allí donde se evidencian las relaciones que existen entre las diferentes tablas, y a la vez determina si el proyecto va a cumplir con su verdadero objetivo.

2.11 Esquemas

Los esquemas en una base de datos, son una forma de organizar la información contenida en la misma. Los usuarios solamente tendrán acceso a aquellos que su rol se lo permita. Dentro de los esquemas se pueden encontrar funciones, operadores y tipos de datos que facilitarán su implementación. En el presente trabajo se definieron dos esquemas:

- Esquema dimensiones: contiene las tablas de las dimensiones generales del AD, y de ellas utilizamos las que se necesiten para implementar el MD.
- Esquema mart_salud: contiene todas las tablas de hechos y las dimensiones propias propuestas en el MD.

2.12 Usuarios y privilegios

Con el objetivo de una mayor seguridad en la base de datos, es necesario definir los usuarios y roles, para con esto, poder realizar su función como trabajador del sistema. Los roles establecidos son:

- Programador de ETL: Su función se basa en la realización de los procesos de ETL en la interacción con la BD Privilegios.
- Administrador: Se le es asignado los privilegios de Select, Insert, Update, Delete, Refresh y Trigger de los datos almacenados en el MD.

2.13 Implementación de los subsistemas de integración

El desarrollo de un software no se puede iniciar hasta no tener bien definida una arquitectura. Esta no es más que un grupo de patrones que sirven de guía para la elaboración del mismo. Para la integración de los datos, la arquitectura queda de la siguiente forma:

- Fuente de Datos: Son archivos de extensión dbf o xls que contienen la información.
- Área temporal: Es el punto intermedio entre la fuente de datos y el MD. Es donde se realiza la integración y transformación de los datos.
- Mercado de Datos: Donde son cargados los datos para su futuro análisis.

2.14 Proceso de Extracción, Transformación y Carga

Extracción de los Datos

Se obtiene toda la información de las distintas fuentes tanto internas como externas. Se cargan los datos de los archivos dbf o excel, para el área temporal y así adaptarlos al modelo relacional que se ha establecido. Estos archivos contienen toda la información referente al MD Salud, la cual será almacenada en las tablas de hecho.

Transformación y Limpieza

Una vez terminado el proceso de extracción, se realiza la limpieza de los datos provenientes de las diferentes fuentes, porque los mismos pueden ser incoherentes, tener errores o estar incompletos.

Con esto se busca obtener datos precisos, completos, y lo más accesibles posibles. 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. La transformación y limpieza es de gran importancia porque en esta etapa es donde se garantiza el resultado final de cómo se van a mostrar los datos, se aplican las reglas del negocio; y se detectan otras posibles deficiencias de la fuente y se corrigen.

Carga

Es donde los datos son cargados al MD, organizados y actualizados, para que puedan ser usados por el cliente de forma satisfactoria. En el presente trabajo se elaboró un área temporal, donde van a ser llevados los datos, y una vez estando allí serán limpiados y transformados para posteriormente ser cargados al MD.

2.15 Implementación de los trabajos

Una vez que la conexión al MD se encuentra en perfecto estado, se procede a la carga del mismo, y el trabajo (job en inglés) es la forma en que se realiza la carga de los datos hacia el MD. Para la correcta realización de un job, se debe tener bien definido cuáles son las dimensiones estáticas y cuáles no, pues en un job solamente se cargan las dimensiones que pueden tomar valores nuevos o cambiar los que tenían anteriormente.

2.16 Implementación del subsistema de visualización de datos. Cubos OLAP

La razón de usar OLAP para las consultas es la rapidez de respuesta y de poder agrupar los datos para garantizar un mejor análisis, obteniéndose los datos más importantes entre toda la información que posee el MD. Para la implementación del módulo de reportes OLAP, es necesaria la creación de los cubos multidimensionales los cuales se realizan utilizando la herramienta Pentaho Schema Workbench, la misma permite generar un fichero de configuración XML. En este fichero de esquema se pueden definir las dimensiones, los niveles de jerarquía de dimensiones, los hechos y conexión con el almacén que sirve los datos para el cubo OLAP.

Se modelaron 41 cubos multidimensionales, en los mismos se especificaron las dimensiones y las medidas. La siguiente imagen muestra el diseño utilizando la herramienta Pentaho Schema Workbench de unos de los cubos modelados para el departamento de Servicios Ambulatorios al modelo de Servicios y equipos rotos a su correspondiente tabla hech_servicios_y_equipos_rotos, el cual está formado por sus dimensiones, y medidas.

Se obtuvo la estructura física del MD, se diseñó el esquema multidimensional con sus respectivos cubos OLAP para agrupar los datos y así facilitar su análisis posteriormente, se identificaron las áreas de análisis, los libros de trabajo y los reportes candidatos, se implementaron y visualizaron los reportes.

3. CONCLUSIONES

1. Mediante la sistematización de la bibliografía accesible se pudo comprobar que existe suficiente fundamentación teórica relacionada con la tecnología de almacén de datos.
2. En el diagnóstico realizado se detectaron problemas en el proceso de toma de decisiones en la dirección de salud y lo que hizo factible la búsqueda de soluciones a partir de la determinación los requisitos funcionales para el desarrollo del almacén de datos.
3. Durante la implementación del almacén de datos se identificaron las metodologías y herramientas que se ajustaban a los requerimientos del producto a elaborar teniendo en cuenta que estas fueran no privativas, las versiones más actualizadas a las que se puede acceder en el momento de la investigación y acorde a los requerimientos técnicos del lugar en que se implantará el sistema.
4. Se validaron las transformaciones a partir de la ejecución correcta de las mismas, se mostraron las vistas y reportes en correspondencia con las necesidades.

4. REFERENCIAS BIBLIOGRÁFICAS

1. Isaith William: Datawarehouse, http://www.sinnexus.com/business_intelligence/datawarehouse.aspx, 2007.

2. Rodríguez Pedro: ORACLE – TecnoXML, http://tecnoxml.wikidot.com/Oracle, 2009.

3. Falcón Yolanda and Leyva Reybaldo: Mercado de Datos Estadístico de Inmigración y Extranjería para el Departamento de Turismo y Comercio de la Oficina Nacional de Estadísticas, La Habana.UCI, 2010.

4. Hernández Asnioby: Documento de Arquitectura del Sistema Almacén de datos para la ONE, La Habana, 2009.

5. Itatí Paola: Tesis de Información, https://docs.google.com/viewer?a=v&q=cache:dCGIuOH-PIJ:200.45.54.90/depar/areas/informatica/SistemasOperativos/PaolaMonog.pdf, 2010.

6. Espinosa Roberto: Pentaho Data Integration, http://churriwifi.wordpress.com/2010/06/01/comparativa-talend-vs-kettle-pdi/, 2010.

7. Curto Josep: Mondrian y su ecosistema, http://www.beyenetwork.es/view/8560, 2008.

5. SÍNTESIS CURRICULARES DE LOS AUTORES

Geidy Acosta Méndez: Estudios Primarios: Antonio Maceo, Cotorro, Habana, Cuba. Estudios Secundarios: Juan Gualberto Gómez, Cotorro, Habana, Cuba. Estudios Medios: Instituto Politécnico de Informática "Gervasio Cabrera", Cotorro, Habana, Cuba. Estudios Superiores: Universidad de las Ciencias Informáticas, Ingeniería en Ciencias Informáticas, 2012. **Actividades Científicas en Proyectos:** 2010- 2011 Analista. Departamento Almacenes de Datos. Facultad Regional "Mártires de Artemisa". Proyecto Almacenes de Datos para la Oficina Nacional de Estadísticas. 2011- 2012 Diseñador de Bases de Datos, Analista. Departamento Almacenes de Datos. Facultad Regional "Mártires de Artemisa". Proyecto Almacenes de Datos para la Administración Provincial de Artemisa.

Disnayle Jorge Chacón: Graduada de la Universidad de las Ciencias Informáticas del año 2010. Especialista general del centro de Ideoinformática. Profesor adjunto a la Facultad 1.