

Lineamientos para la Selección del Enfoque Inmon o Kimball en Proyectos de Data Warehouse

Flor Hernández Pérez / Ingeniera de Sistemas
Especialista en Desarrollo de Soluciones Informáticas
Docente Unicomfacauca
fhernandez@unicomfacauca.edu.co

Resumen: El Data Warehouse (DW) es una herramienta tecnológica e importante para organizaciones en general, este artículo describe algunos casos en los cuales se han implementado herramientas de este tipo. Para la implementación y desarrollo de este tipo de herramientas se consideran dos estilos: el Estilo de Inmon y el Estilo de Kimball, pero el problema no radica en el desarrollo y la implementación de la misma sino en el dilema de seleccionar uno de los dos estilos pioneros en esta temática. Lo que se pretende en este artículo es dar a conocer los avances respecto a este tema para en el futuro dar a conocer los lineamientos necesarios para realizar la selección correcta del estilo para diseño de un DW.

Palabras clave: Data Warehouse, lineamientos, teoría Gestáltica, teoría Semántica de Red.

Introducción

Desde el inicio de la era de la computadora en general todas las organizaciones se ha preocupado por dar un mejor manejo a los datos que éstas generan, buscando mejorar la calidad de datos entregados a los usuarios. Estas organizaciones han manipulado sus datos utilizando aplicaciones operacionales y bases de datos. Con el paso del tiempo las organizaciones han acumulado información valiosa generada por la misma, pero muchas de las organizaciones no poseen sistemas o la tecnología para encontrar esta información y aprovecharla de forma estratégica.

Un Data Warehouse es una herramienta que apoya la toma de decisiones y actualmente muchas organizaciones se han inclinado por hacer parte del grupo de empresas con esta nueva herramienta tecnológica que hará que su información sea utilizada de una forma óptima en diversidad de aplicaciones operacionales. [1].

Los sistemas de información antes de ser implementados dentro de una organización deben ser diseñados utilizando diferentes métodos o modelos, estos modelos poseen cierta cantidad de características y pasos a seguir, todos con la finalidad de entregar un modelo lo más cercano posible y entendible para los usuarios y haciendo que la

información que contenga sea de calidad y eficiente para el apoyo en la toma de decisiones.

Dentro de los Modelos para el diseño de un Data Warehouse se puede hablar de Modelo Dimensional de Ralph Kimball (SSD Diagram Star-Schema) y el Modelo Relacional de William Inmon (ERD Diagram Entity-Relationship). Estos modelos también son conocidos como modelo Estrella y modelo Copo de Nieve.

La decisión de construir un Data Warehouse que utiliza un Modelo Relacional (ER) o un Modelo Dimensional (DM) ha creado una controversia, en algunos casos, el Modelo Relacional de Inmon es preferido [2]. Pero en la mayoría existe inclinación hacia el Modelo dimensional de Kimball ya que esta técnica es considerada como la más apropiada para el apoyo en la toma de decisiones.

Marco Teórico

En el manifiesto de modelo dimensional [1], se muestran las diferencias entre estos dos modelos al momento de la construcción de un Data Warehouse.

El modelo ER evita la redundancia de datos, pero éste factor no garantiza la eficiencia en la manipulación de

los mismos e incrementando la incoherencia. Evitar la redundancia de datos es importante para el procesamiento de transacciones debido a la simplicidad y que son determinísticas. La disciplina del modelo ER garantiza el éxito del procesamiento de transacciones, pero no garantiza que el usuario pueda consultar la base de datos.

La cantidad de entidades en un modelo ER se torna en una situación incómoda porque es difícil para el usuario entender y recordar el modelo, se dice que esto se debe a la falta de una interfaz grafica (GUI) que sea utilizable para los usuarios. No resulta útil realizar consultas en el software del un ER, debido a la optimización de las mismas llevando a tomar decisiones equivocadas y en consecuencia desastrosas.

El Modelo Dimensional (DM) es una técnica de diseño lógico que busca presentar los datos en una norma, marco intuitivo que permite el acceso de alto rendimiento. Es intrínsecamente dimensional y se adhiere a una disciplina que utiliza el modelo relacional con algunas restricciones importantes. Cada modelo dimensional se compone de una tabla con una clave multiparte, denominada tabla de hechos, y una serie de pequeñas tablas llamadas tablas dimensión. Cada dimensión tiene una única clave principal que corresponde exactamente a uno de los componentes de la clave en la tabla de hechos. [3].

El modelo ER (Figura 1), muestra una visión detallada de todas las entidades que conforman el modelo, donde cada casilla del diagrama puede contener muchas entidades más. El diseño de DM equivalente aislaría cada proceso de negocio y lo rodearía solamente por sus dimensiones relevantes.

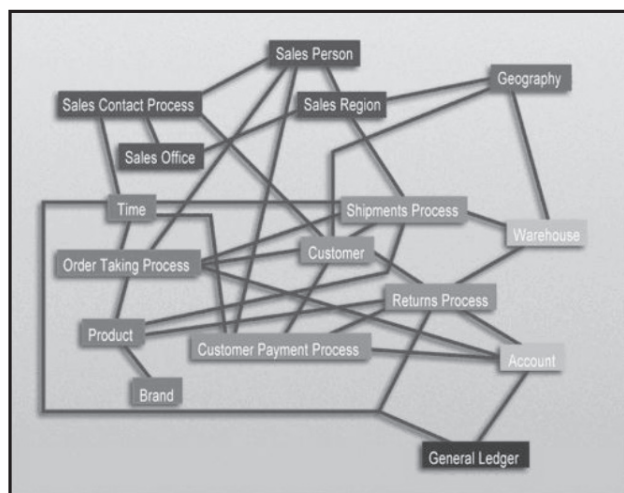


Figura 1. Modelo Entidad/Relación (ER)

En la Figura 2 se presenta un modelo dimensional detallado de los puntos de venta al por menor, donde se diferencia fácilmente la tabla de hechos con sus medidas y las tablas de dimensiones asociadas a esta tabla de hechos.

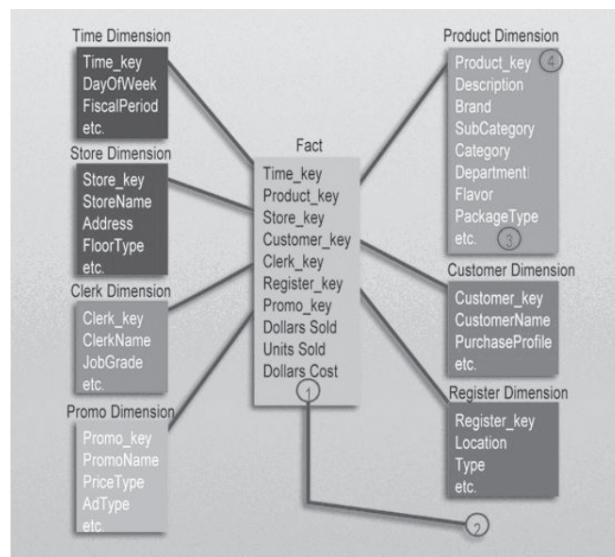


Figura 2. Modelo dimensional (DM)

Teniendo en cuenta las figuras anteriores se puede empezar notando la diferencia que existe entre cada uno de los modelos, en muchos casos se piensa que el modelo DM no posee toda la información para entregar datos de calidad al usuario final, debido a la sencillez de su diagrama.

Es cierto que existe una relación entre los dos modelos y precisamente esa relación consta de que considerando la extensión del modelo ER, el DM se convierte en una versión mucho más sofisticada rompiendo un diagrama ER en varios diagramas DM, disminuyendo la complejidad del modelo inicial.

Existen ciertas características de DM que hacen de este modelo en muchas ocasiones preferido para iniciar con el diseño de un Data Warehouse, entre estas características se puede nombrar su marco normativo y la previsibilidad, con interfaces de usuario más comprensibles haciendo que el procesamiento y análisis sea más eficiente.

Durante el estudio y análisis de este artículo se observó que el modelo DM apunta a entregar información de calidad, cumpliendo con requisitos que se consideran indispensables para el usuario final.

Estado Del Arte

Una comparación entre el diagrama estrella y diagrama ER[4] se basa en un experimento realizado con personas especializadas en ER y personas que desconocen los modelos de datos. En este artículo se presenta una inclinación hacia el DM, mostrando que es mucho más sencillo de entender y recordar por los usuarios, debido a que un usuario al observar un diseño de un DW utilizando DM lo entienda más rápidamente y al realizar las consultas la información entregada será de calidad sin depender tanto del profesional en informática para tener acceso a los datos. Si un usuario final entiende más rápido un DM las preguntas realizadas al Data Warehouse son mejores.

Otro punto importante a considerar es que los expertos en ER respondieron que de todos los ejercicios realizados, las entidades que más recuerdan son las que fueron construidas con el DM.

La teoría semántica de red y el comportamiento humano de la memoria es utilizada para explicar el por qué tanto personas experimentadas como inexpertos recuerdan con facilidad y comprenden mejor en DM llega a los siguientes resultados. Estos investigadores han llegado a la conclusión de que el hombre guarda información organizada y ésta es la razón por la cual se recuerda el DM.

Otra teoría incluida en el artículo es la teoría gestáltica que se encarga de proporcionar aspectos visuales y tiene implicaciones significativas para la investigación mental. El desarrollo de esta hipótesis arroja como resultado que el DM es más fácil de recordar que el ER, porque organizar las entidades de dimensión alrededor de la tabla de hechos da una pauta significativa (por tanto una estructura más fácilmente recordada de almacenamiento).

DM también se adhiere a los principios gestálticos de continuidad (sucede cuando las líneas entre entidades implican una relación directa entre la dimensión y el hecho) y proximidad (sucede cuando la relación física entre dimensiones alrededor de la tabla de hechos implica relaciones próximas entre esos elementos en el esquema).

Según [5], el ciclo de vida de un Data Warehouse se ha convertido en un dilema al seleccionar el Esquema de diseño de este tipo de herramienta tecnológica con soporte en la toma de decisiones. Diversidad de compañías de venta al de tal en Estados Unidos han experimentado con

variedad de herramientas para soporte de negocios que no ha arrojado resultados óptimos. El acceder a los datos cuando se han utilizado estas herramientas no ha sido fácil, ya que la información no concuerda con la aplicación del sistema. Estas aplicaciones guardan una pequeña historia de los mismos, discerniendo el conocimiento y haciendo que la producción del negocio sea pequeña. Debido a esto la compañía de venta al de tal en Estados Unidos decide en 1995 usar Data Warehouse y dejar en libertad los datos.

A la hora de iniciar un proyecto de DW, se considera que lo difícil no es realizar las preguntas al DW sino que lo difícil es seleccionar el esquema de diseño. En este artículo se dice que los esquemas difieren de la disposición de los datos dentro del DW.

En este artículo se presenta un proyecto realizado en una empresa de venta al por menor que se desarrolló utilizando el esquema de Inmon, para la selección de este esquema se tuvo en cuenta factores como: conocimiento de los empleados de la tecnología a usar, preferencias en el nivel detalle de los datos, recomendaciones de los usuarios, presupuesto para el proyecto, tecnologías existentes y la percepción de ventajas. Como se aprecia el estilo fue seleccionado teniendo en cuenta las circunstancias del entorno empresarial y técnico.

El bajo nivel en que suelen mantenerse los datos en este estilo es un factor de detalle y eso es lo que los usuarios necesitaban. Fue necesaria una fase para el suministro de datos (la normalización de los mismos), mientras que con el enfoque de Kimball era preciso realizar una fase más para el suministro de datos, para que los datos sean denormalizados.

Otro aspecto a tener en cuenta fue la experiencia de la comunidad empresarial para apoyar la autosuficiencia del usuario final.

La capacidad del estilo Inmon de crear los datos dependientes entre los data mart fue un factor crítico, debido a que en algunas situaciones se requiere informes repetitivos o la necesidad específica de la aplicación.

Hay muchos debates filosóficos, los obstáculos, los pro y los contra en cuanto a la selección de un método de DW. Sin embargo sea cual sea la metodología que se elija debe cumplir con los requerimientos de negocio de la organización y ser flexible y escalable. En otras palabras, la metodología debe reducir al mínimo las diferencias

entre los procesos de negocio y la tecnología que se utiliza para dirigir la empresa.

Debido a la necesidad de detalle que era requerido en esta empresa de venta al por menor, el estilo de Inmon se acogió de una manera excelente entregando resultados de éxito para la empresa, por lo tanto el aporte que esta investigación presenta se centra en analizar las necesidades de la empresa y la tecnología que hay en ella y para ello se hacen una serie de recomendaciones que entregaron buenos resultados [5].

El artículo "MODELADO DIMENSIONAL DE DATOS" se basa en un estudio donde se muestra que el modelado dimensional de datos se puede aplicar dependiendo de la tarea que se especifique, es decir sin cumplir a cabalidad con las normas que necesita el modelo para entregar al usuario datos de calidad, esto en algunos casos ha dado buenos resultados pero concluye que lo mejor que se puede hacer es que teniendo en cuenta la cantidad de tiempo que lleva este estilo dando buenos resultados y hoy en día podemos efectuar un mayor análisis de información es gracias a su correcta aplicación. Por lo tanto dentro de la investigación que se está iniciando se debe tener en cuenta que si el estilo fue creado con ciertos parámetros éstos deben respetarse y utilizar correctamente porque de esto depende la efectividad del modelo y por tanto el éxito de un Data Warehouse.

Otro de los artículos que se ha incluido en la investigación es "BI INTELIGENCIA DE NEGOCIO", este artículo da a conocer variedad de conceptos que permiten entender a un nivel más profundo el modelado de un Data Warehouse utilizando BI inteligencia en el negocio. El objetivo de este artículo es dar una introducción al concepto de BI y a algunas de sus fases a través de un proyecto de este tipo en el que se optó por las técnicas OLAP (Online Analytical Processing) para el análisis de la información [7].

Conclusiones

El aporte que da a la investigación no es tan grande pero hay ciertas pautas que se cree son importantes a la hora de entrar en un nivel más profundo en el desarrollo del objetivo de la investigación. En este artículo plantean un ejemplo donde explican de forma explícita y general como debe hacerse la preparación de los datos, el modelado y la evaluación del ejemplo.

Dentro de este artículo se definen los datos como la materia prima, por lo tanto a la hora de realizar la recolección

de los mismos es importante saber organizarlos porque al momento que el usuario les atribuye un significado especial pasan a convertirse en información. Este tipo de aporte puede ser relevante, pues no se está discutiendo cómo debe ser el modelado de un data Warehouse pero en cierto caso puede ser importante tenerlo en cuenta y observar los resultados que arrojó el ejemplo.

En cuanto a la preparación de los datos, es una colección de procesos que limpian, transforman, combinan, y preparan los datos originales para su utilización en el data Warehouse.

En este punto hay que asegurarse de unas cuantas cosas:

- Que los datos tengan la calidad suficiente: es decir, que no contengan errores, redundancias o que presenten otro tipo de problemas.
- Que los datos sean los necesarios, quizás algunos no nos harán falta y quizás tendremos que añadir otros.
- Que estén en la forma adecuada: muchos métodos de construcción de modelos requieren que los datos estén en un formato determinado que no ha de coincidir necesariamente con el que están almacenados [6].

Estos son aspectos importantes a la hora de seleccionar un modelo, puesto que si los modelos cuentan con datos efectivos el resultado de elegir uno u otro no será el más óptimo.

También se puede hablar de cierta ventaja que tiene el modelado multidimensional, la cual es que facilita los análisis complejos al estar muy próximos a la manera de pensar del analista, y la visualización de los datos en el datamart para procesos de toma de decisiones, reduciendo la confusión y disminuyendo las interpretaciones erróneas.

Además, ya que los datos están almacenados físicamente en una estructura multidimensional o base de datos n-dimensional, la velocidad de estas operaciones es varias veces superior y más consistente de lo que es posible en otras estructuras de bases de datos. La combinación de simplicidad y velocidad es uno de los principales beneficios del análisis multidimensional.

Este se puede considerar como un punto más a favor del modelado multidimensional que debe tenerse en cuenta dentro de la investigación, claro que se debe tener en

cuenta que se deberá realizar un análisis más profundo para evitar posibles confusiones y así evitar información errada o quizá estar adentrándose en terrenos que no están incluidos en la investigación.

Por medio del ejemplo realizado en el artículo se confirma de forma general algunos aspectos que se nombraron en el análisis del segundo artículo estudiado, como son poca dependencia de profesional de informática y que el usuario cuente con una herramienta flexible para acceder a la información.

El propósito final de la investigación es proveer o hallar unos lineamientos con las bases suficientes para orientar a una organización que desee emprender un proyecto de Data Warehouse, en este punto de la investigación puede decirse que son muchos los puntos a favor de DM.

Referencias

- [1] Cacerea.C.,<http://www.programacion.com/bbdd/tutorial/warehouse/> . 1998-2003.
- [2] W.H. Inmon, C. Imhoff, R. Sousa, Corporate Information Factory, Wiley, New York, 1998.
- [3] R. Kimball, A dimensional modeling manifesto, DBMSmag.-com, <http://www.dbmsmag.com/9708d15.html> (1997).
- [4] Corral, Karen. David Schuff., D. Roberth. Lois St. "The Impact of Alternative Diagram on the Accuracy of Recall: a Comparison of Star-Schema Diagrams and Entity-Relationship Diagram". Decision Support Systems 42 (2006).
- [5] Lawyer J.,Chowdhury S., "Best Practices in Data Warehousing to Support Business Initiatives and Needs" Proceedings of the 37th Hawaii International Conference on System Sciences - 2004
- [6] Verastegui H. C., "MODELADO DIMENSIONAL DE DATOS" Universidad de los andes Bogotá Colombia.(2006).
- [7] Llombart O., (oalonso@uoc.edu) BI INTELIGENCIA DE NEGOCIOS CMS-SPAIN.COM