Metodología crisp para la implementación Data Warehouse

Methodology crisp for data warehouse implementation

Octavio J. Salcedo Parra
Ingeniero de Sistemas y Magister en Teleinformatica y en Economia. Docente de la Universidad Distrital Francisco José de Caldas. Bogotá, Colombia. ojsalcedop@udistrital.edu.co

Rita Milena Galeano
Ingeniera de Sistemas. Consultora de Oracle. Bogotá, Colombia. ritamilena84@hotmail.com

Luis G. Rodriguez B.
Ingeniero de Sistemas. Docente del SENA. Bogotá, Colombia. lgrodriguezb@udistrital.edu.co

Clasificación del artículo: investigación (Conciencias)

Fecha de recepción: mayo 21 de 2009 Fecha de aceptación: noviembre 17 de 2009

Palabras clave: Bases de datos, Cubos, Dimensiones, Estructura en estrella, Metadatos, Subsistemas, Tiempo real.

Key words: Databases, Cubes, Dimensions, Star structure, Metadata, Subsystems, Real time.

Resumen

En la actualidad la generación de informes claros, concisos y ante todo veraces, con base en la información de las empresas, es un elemento fundamental en la toma de decisiones. Debido a esta necesidad inminente surge Data Warehouse como recurso esencial para la realización de dicho proceso, cimentado primordialmente bajo la filosofía OLAP y el cual utiliza los concepto EIS y DSS para la realización de los informes. Dentro de los procesos que se realizan para la construcción de las bodegas de datos se destacan principalmente la extracción, transformación y manipulación de la información para la posterior definición de los metadatos, los cuales son utilizados para la definición del Data Warehouse como sistema integrado. La tendencia hacia la que apunta la "inteligencia de negocios" es la divulgación de la información, tanto a nivel gerencial como a todo aquel que la necesite desde diferentes dimensiones y niveles asociados, para lograr obtener informes consolidados o detallados que faciliten la síntesis de determinado proceso empresarial y que repercutan directamente en la toma de decisiones, objeto que en últimas constituye el objetivo mismo de los Data Warehouses.

Para llevar a cabo la implementación de dicho proceso es necesario disponer de una metodología adecuada, de tal manera que el proyecto se diseñe bajo la estructuración de unos estándares internacionales, los cuales constituyen los cimientos para la obtención de excelentes resultados sobre la puesta en marcha del proyecto.

Abstract
Currently, the generation of crystal clear reports, concise and above all based on true corporate information is a fundamental element in decision making, because this imminent need arises data warehouse as an essential resource for conducting the process, primarily founded on the philosophy using the concept OLAP and EIS and DSS for the completion of reports. Within the processes carried out for construction of the data warehouse is mainly involving the extraction, processing and handling Information for further definition of the metadata which in turn are used to define the data warehouse as an integrated system. The trend towards pointing BI, is to the dissemination of information both management and to all who need it from different dimensions and levels associated in order to obtain consolidated or detailed reports to facilitate the synthesis of certain business process that directly impact the decision-making, which at last is the same purpose of the data warehouse.

To carry out the implementation of this process is necessary to have an appropriate methodology, so that the project was designed under the structure of international standards, which are the foundation for obtaining excellent results on project implementation.

* * *

1. Introducción

Uno de los principales fines de la computación desde sus orígenes ha sido presentar una herramienta de apoyo al hombre, de tal forma que disminuya sus cargas mecánicas y repetitivas, y aumentar de esta manera su nivel de vida en cuanto al ahorro de tiempo para redistribuirlo en actividades de mayor importancia para él mismo, e influyendo significativamente en la innovación y evolución tecnológica.

En consecuencia de ello, el manejo de los datos se convirtió en un elemento clave para la extracción de información, de tal forma que el individuo pueda mantenerse enterado de la situación que éstos presenten. Como resultado, día a día, la información se ha convertido en un recurso muy valioso y desde hace tiempo atrás se usan las bases de datos como mecanismo para guardar la información dentro de su arquitectura.

Es por ello que la tendencia futura del manejo de información está orientada y focalizada principalmente hacia la inteligencia de negocios y precisamente temas como este, que anteriormente muy poco se trataba y se desconocía gran parte del mismo, ahora está en las mentes de todos y tomando cada vez más fuerza.

En la actualidad muchas casas de *software* están agregando dentro de sus servicios la construcción de Data Warehouse como herramienta ventajosa en el proceso de toma de decisiones, pues además de proveer recursos valiosos, resulta ser una pieza clave para el *puzzle* del *Business Intelligence*.

Es de esta manera como se introduce el concepto de Data Wareouse como elemento sólido y robusto en los diversos procesos empresariales. Cabe rescatar que el éxito de la inteligencia de negocios, específicamente en los Data Warehouses como procesos y no como productos reside sobre la metodología implementada y las técnicas aplicadas para la realización del proyecto.

2. Generalidades

Hoy en día las organizaciones orientan sus mayores esfuerzos hacia la maximización de sus ingresos, y por ende minimización de los costos, en función de la productividad empresarial. Para conseguir dichos

resultados se requiere una serie de elementos que se deben ver reflejados en la estructura piramidal de la empresa, es decir, desde los empleados de más bajo nivel hasta los altos ejecutivos.

Para lograr un mayor rendimiento se debe realizar una serie de procesos y análisis estratégicos que involucren las variables de la empresa y su entorno. Dichos procesos requerirían de un esfuerzo adicional por parte de la organización para lograr sus objetivos. Es por ello que la tecnología Data Warehouse resulta de vital importancia en este significativo peldaño, pues posee elementos robustos que proporcionan confiabilidad y veracidad en la toma de decisiones.

Los Data Warehouse pueden estar compuestos por Data Marts, que son una particularización de las bodegas de datos, que heredan las mismas características de los Data Warehouse y cuyo enfoque principal recae sobre los departamentos, áreas o módulos específicos de la empresa. Además de ello, podemos precisar los Data Marts como partes indispensables y necesarias para integrar el sistema, y que además influyen en el manejo de un control más adecuado de los datos bajo la filosofía OLAP (On-Line Analytical Processing)[1] que a su vez usa estructuras multidimensionales para planear una mejor distribución del modelo y para que mucho más confiable, efectivo y transparente para el usuario final o persona que requiera del análisis de los datos.

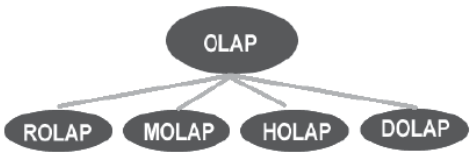

Figura 1. Subdivisión OLAP.

[1] OLAP: Análisis en línea de los datos que permite utilizar fórmulas matemáticas y análisis estadístico para consolidar y resumir los datos procesados. Disponible en: http://www.monografias.com/trabajos16/datawarehouse/datawarehouse.shtml

3. Conceptos asociados

La importancia que ha adquirido la tenencia de la información como recurso potencial de las empresas ha desatado una verdadera revolución tecnológica en torno de esta. Es por ello que día a día se desarrollan nuevos mecanismos que facilitan la interpretación, análisis y utilización de la información como herramienta clave para atacar problemas, brindar soluciones y realizar proyecciones para ofrecer una mejor gestión en términos de eficiencia y facilidad de uso para el beneficiario.

Data Warehouse nace como tentativa a la construcción de un nuevo concepto tecnológico y herramienta competitiva que promete diseñar nuevas alternativas de negocio con base en la información de la empresa, generando informes consolidados o detallados según los niveles definidos en las diferentes dimensiones; además, con la particularidad de ofrecer información de carácter gerencial para la toma de decisiones.

W.H. Inmon, considerado el padre de las bodegas de datos en el 92, define los Data Warehouse como: "Un sistema orientado al usuario final, integrado, con variaciones de tiempo y sobre todo una colección de datos como soporte al proceso de toma de decisiones". Por otra parte, Ralph Kimball, considerado como uno de los más importantes precursores y padre del concepto Data Warehouse, lo define como: "una copia de los datos de la transacción estructurados específicamente para preguntar y divulgar" [5].

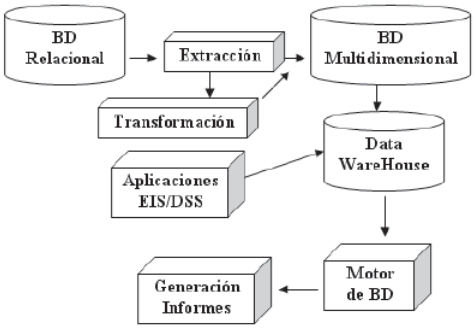

Figura 2. Proceso Data Warehouse.

Esta última definición resulta muy interesante y completa, pues corresponde a la síntesis del proceso como tal ya que a grandes rasgos lo podemos proyectar de la siguiente manera: abarca desde la replicación de la información de la BD relacional hacia la BD multidimensional, pasando por el proceso de extracción y transformación de la información para obtener así los datos en la forma requerida y que en últimas conlleven a la generación de los informes que deberán contener la información estructurada y facultada para ayudar al usuario final en el proceso de la toma de decisiones.

Este proceso puede ser visto de dos formas, una global que aplicaría a Data Warehouse, y una parcial que vendría a estar representada por los Data Marts, donde estos últimos serían los subsistemas que realizarían funciones específicas, coordinadas y correlacionadas para conformar el todo o el sistema Data Warehouse. Esto quiere decir que en un modelo multidimensional es implementando un Data Mart con todas las implicaciones a las que se acoge la tecnología Data Warehouse, ya que los Data Marts son subsistemas que heredan todas las características de sus padres y a su vez pueden conformar un sistema mismo.

Según Ralph Kimball, un Data Mart es: "Un subconjunto lógico del Data Warehouse completo*"*. A esta definición la Global Communications agrega en enero de 1999: "Es un almacén de datos diseñado para dar soporte a un departamento o unidad de negocio en particular. Puede ser independiente, parte de una red distribuida de Data Mart, o dependiente de los datos en un almacén Central (DW)" [7].

Figura 3. Data Marts.

En este orden de ideas el Data Mart es un subsistema dentro del sistema DW que hereda todas las propiedades y cumple funciones específicas sobre un determinado módulo, departamento o área de la cual forma parte el propio DW.

Para cumplir con los objetivos, la tecnología Data Warehouse la cual es manejada principalmente en el área de inteligencia de negocios y cuya razón principal para ser utilizada es por la ventaja en tiempo de respuesta a la hora de realizar consultas utiliza una serie de mecanismos en cuanto a la estructura interna del manejo lógico de la BD multidimensional, en términos de reorganización y acomodación de los datos en cubos multidimensionales como estructura sencilla, pero lo suficientemente sólida como para proporcionar los beneficios antes mencionados.

Adicional a ello, y con una relación intrínseca en todo el proceso, el conjunto de técnicas que comprende el Data Warehouse maneja los conceptos EIS (Sistemas de Información Ejecutiva) y DSS (Sistemas de Soporte en la toma de Decisiones) como herramientas claves y orientadas fundamentalmente a los usuarios finales, ya que son estos quienes visualizan los reportes, y es precisamente hacia ellos donde las herramientas EIS y DSS realizan su mejor trabajo y definen las pautas para mostrar los resultados sin que haya ningún tipo de afectación hacia los datos. Por el contrario, la información retornada en dichos informes realizados bajo estos conceptos es la que en últimas facilita el trabajo de la toma de decisiones.

La tecnología Data Warehouse está sustentada principalmente en dos grandes sistemas: el primero es el sistema técnico–operacional, encargado de las tareas principales de la empresa, y el segundo es el sistema de soporte de decisiones, cuyo fin principal está orientado hacia un "planeamiento, previsión y administración de la organización"[2], Es precisamente

[2] Data Warehousing. Disponible en: http://www.sqlmax.com/dataw1.asp

con base a estos dos grandes sistemas donde se encuentra reflejado el proceso Data Warehouse.

Los elementos que caracterizan un Data Warehouse son:

3.1. Particularización de las necesidades del cliente

Se enfoca principalmente en los datos propios de cada ente, y descarta así la información innecesaria para la realización del proceso, asegurando de esta manera la personalización del sistema y la solución óptima en el ámbito de la toma de decisiones.

3.2. Unificación

Independientemente de las diferentes formas en la que se encuentren almacenados los datos en la BD origen, al ser llevados al DW por medio del proceso de transformación deben coincidir en su estructura, medida y forma en general; de esta manera, el proceso de análisis de la información resulta menos engorroso y útil en términos generales para el analista.

3.3. De tiempo variable

Hace alusión a la relación histórica de los datos, es decir, pueden manejar una línea de tiempo que oscila aproximadamente entre cinco y diez años, y dichos datos no pueden ser alterados una vez alojados en el DW.

3.4. No volátil

La estabilidad de la información en términos generales, la persistencia de los datos y la conservación en el tiempo es lo que precisa la robustez del Data Warehouse.

En ese orden de ideas, las operaciones que se realizan en el entorno DW son extracción de la información, transformación de los datos para lograr la unificación, carga en la arquitectura DW y consulta.

Otro concepto importante que se maneja entorno a esta nueva tecnología es el de los metadatos, sobre los cuales están cimentados todos los conceptos del DW, ya que son datos que definen datos y resultan variantes dependiendo del ambiente en el cual se encuentre; es decir, podemos establecer metadatos dependiendo de la actividad económica de la organización. Además, corresponden a la documen tación del proyecto y son los cimientos para la base del conocimiento.

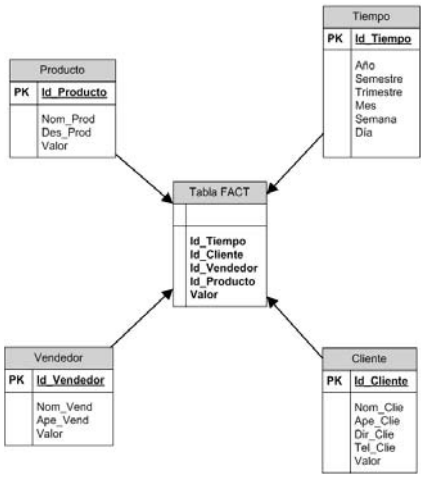

Figura 4. Estructura *Snowflake*.

Cabe resaltar que la tecnología Data Warehouse maneja una estructura en estrella que a su vez se convierte en una estructura de copo de nieve, dependiendo del nivel de detalle al cual se quiera llegar en el diseño de los cubos multidimensionales, los cuales son arreglos lógicos donde cada eje es una dimensión y cada punto dentro del cubo es la intersección de las dimensiones establecidas, elemento que permite retroalimentar una proyección de los datos que nos interesan vs. tiempo.

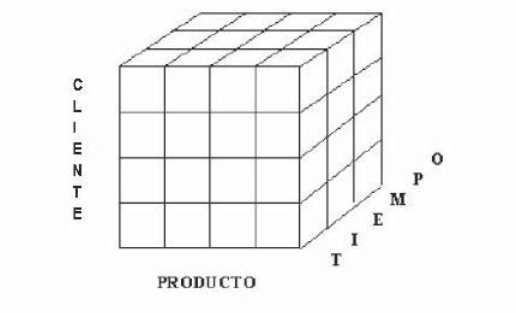

Figura 5. Cubo multidimensional.

Los beneficios que afloran con la implementación de inteligencia de negocios que llevan implícito el concepto de Data Warehouse son muchos y el alcance depende de los objetivos que se planteen y de la trascendencia que adquieran los datos para el análisis de la información; lo importante siempre es la definición inicial de los requerimientos y hacia dónde va encaminado el proyecto. Esos son los elementos claves para la creación de un correcto diseño y posterior implementación del Data Warehouse robusto, sólido, eficiente, confiable y ante todo útil para el manejo de la toma de decisiones. El mapa de Data Warehouse está basado en la arquitectura SQL Server.

4. Metodología

CRISP-DM (*Cross- Industry Standard Processfor Data Mining*). La metodología CRISP es una de las principales metodologías por seguir por los analistas en la inteligencia de negocios, donde se puede rescatar primordialmente Data Warehouse y Data Mining.

La metodología CRISP está sustentada en estándares internacionales que reflejan la robustez de sus procesos y que facilitan la unificación de sus fases en una estructura confiable y amigable para el usuario. Además de ello, esta tecnología interrelaciona las diferentes fases del proceso entre sí, de tal manera que se consolida un proceso iterativo y recíproco. Otro aspecto fundamental de esta tecnología es que es planteada como una metodología imparcial o "neutra respecto a la herramienta que se utilice para el desarrollo del proyecto de Data Warehouse o Data Mining siendo su distribución libre y gratuita" [8].

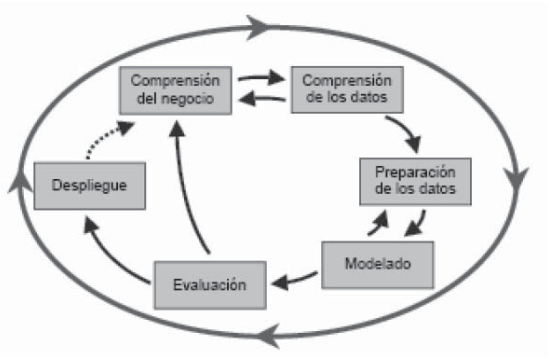

Figura 6. Ciclo CRISP.
Tomada de http://www.dsic.upv.es/~jorallo/master/dm5.pdf

El ciclo de vida del proyecto según la metodología CRISP está basado en seis fases cambiantes entre sí y nunca terminantes, lo cual lo postula como un ciclo en constante movimiento.

4.1. Comprensión del negocio

Se trata de entender claramente los requerimientos y objetivos del proyecto siempre desde una visión de negocio. Esta fase se subdivide a su vez en las siguientes categorías:

- Definición de los objetivos de negocio (inicial, objetivos de negocio y criterios de éxito del negocio).
- Evaluación de la situación (inventario de recursos, requisitos supuestos y requerimientos, riesgos y contingencias, terminología y costes y beneficios).
- Definición de los objetivos del Data Warehouse (objetivos y criterios de éxito).
- Realización del plan del proyecto (plan del proyecto y valoración inicial de herramientas y técnicas).

4.2. Comprensión de los datos

Es conseguir y habituarse con los datos, reconocer las dificultades en la calidad de los datos y reconocer también las fortalezas de estos mismos que pueden servir en el proceso de análisis. Sus subdivisiones son:

- Recolección inicial de datos (informe de recolección).
- Descubrimiento de los datos (informe descriptivo de los datos).
- Exploración de los datos (informe de exploración de los datos).
- Verificación de calidad de los datos (informe de calidad).

4.3. Preparación de los datos

Es analizar los datos realmente importantes en el proceso de selección, depuración y transformación. Sus subdivisiones son:

- Selección de los datos (motivos para incluirlos o excluirlos).
- Depuración de los datos (reporte de depuración).
- Estructuración de los datos (generación de atributos y registros)
- Integración de los datos (agrupar los datos).
- Formateo de datos (informe de la calidad de datos formateados).

4.4. Modelado

Es la aplicación de técnicas de modelado o de Data Warehouse. Sus subdivisiones son:

- Selección de la técnica de modelado (técnica y sus supuestos).
- Generar el plan de pruebas (plan de pruebas).
- Construcción del modelo (parámetros escogidos, modelos, descripción de los modelos).
- Evaluación del modelo (evaluar el modelo, revisión de los parámetro elegidos).

4.5. Evaluación

Esta fase es muy importante y decisiva, pues corresponde a la evaluación de la escogencia de los modelos anteriores y la toma de decisión respecto a si realmente son útiles en el proceso. Sus subdivisiones son:

- Evaluar resultados (valoración de los resultados respecto al éxito del negocio, modelos aprobados).
- Proceso de revisión (revisar el proceso).
- Determinación de los pasos siguientes (listado de posibles acciones, técnica modelada).

4.6. Despliegue o divulgación

Es la fase de implementación o de divulgación de los modelos anteriormente escogidos y evaluados. Sus subdivisiones son:

- Plan de divulgación o implementación (plan de implementación).
- Plan de monitoreo y mantenimiento (plan de monitoreo y mantenimiento).
- Presentación del informe final (informe final, presentación final).

- Revisión del proyecto (documentación de la experiencia).

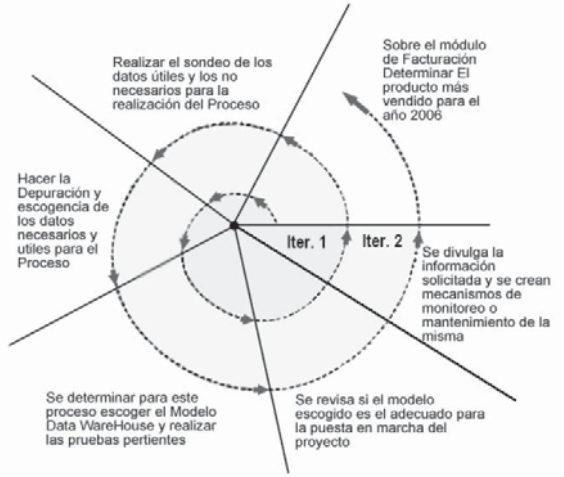

Figura 7. Proceso CRISP.

5. Resultados

5.1. Lenguajes de consulta inductivos

Los lenguajes de consulta inductiva son más que una simple consulta; se proyectan hasta los límites de la búsqueda de patrones, los cuales ceden a los usuarios los privilegios de restringir la búsqueda inductiva bajo las siguientes consideraciones:

"La parte de la base de datos a ser minada (también llamada la vista minable o vista relevante). El tipo de patrón/reglas a ser minado (también llamado restricciones del conocimiento). Cuantificadores estadísticos: representatividad (support) %, precisión (confidence/accuracy) %. Otras propiedades que el patrón debería cumplir (número y forma de las reglas, interés, novedad, etc.)"[3] .

[3] http://www.dsic.upv.es/~jorallo/master/dm5.pdf

Propuesta M-SQL. Basada en modelos de consulta [10]

Ejemplo:

SELECT FROM MINE (T): R

WHERE R. Consequent = {(Age = *)}

R. Support> 1000

R. Confidence> 0.65;

R es una variable de regla y se puede utilizar:

R. Consequent

R. Body (antecedente)

R. Support

R. Confidence

Propuesta DMQ (Data-Mining Query) language (Ng et al.1998)

- Utiliza la sintaxis del SQL para la vista minable.
- También basado en modelos de consulta.

Ejemplo:

Esquema:

SALES (customer_name, item_name,transaction_id)

LIVES (customer name, district, city)

ITEM (item name, category, price)

TRANSACTION (transaction id, day, month, day)

Consulta Inductiva (lenguaje natural):

"Buscar las ventas de qué artículos baratos (con una suma de precios menor que $100) que puede

motivar las ventas de qué artículos caros (con el precio mínimo de $500) de la misma categoría de los clientes de Vancouver [12].

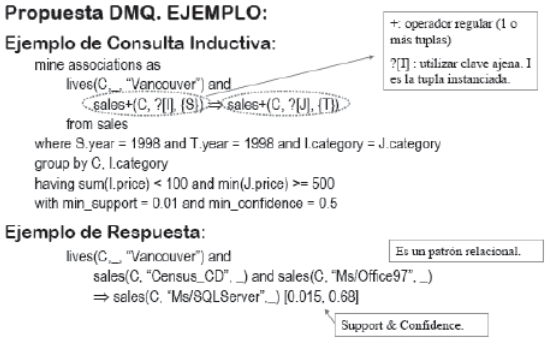

Figura 8. Consulta inductiva.

Los lenguajes de consulta inductiva, de extensión del protocolo de acceso a BB.DD. OLE DB implementan una extensión de SQL que trabaja con DMM (Data Mining Model) y a su vez permiten:

- Crear el modelo.
- Entrenar el modelo.
- Realizar predicciones.

La creación de un modelo (DMM) se puede estructurar de la siguiente manera:

```
CREATE MINING MODEL RiesgoDeCredito
(
[ID]                LONG   KEY,
[PROFESION]         TEXT   DISCRETE,
[DEUDA]             TEXT   DISCRETE,
[EDAD MORA]         LONG   CONTINUOS,
[NIVEL_RIESGO]      TEXT   DISCRETEPREDICT
)
USING [DecisionTree]
```

Para este nuevo modelo de minería de datos se ha logrado crear un modelo vacío que deberá ser alimentado por sentencias especiales del modelo DMM. Para el caso particular de la creación del modelo, las cuatro primeras definiciones (KEY, DISCRETE, DISCRETE, CONTINUOS) son los atributos de entrada del sistema, y el atributo (DISCRETEPREDICT) es el atributo de salida del sistema, donde se mide el nivel de riesgo para cada uno de los personajes identificados luego de realizado el proceso. Por último, [DecisionTree] es el tipo de modelo específico.

Para alimentar el modelo es necesario utilizar la cláusula INSERTINTO, utilizada comúnmente en bases de datos, pero que en minería de datos requiere de una sintaxis y procesos específicos, tales como:

```
INSERTINTO [RiesgoDeCredito]

([ID], [PROFESION], [DEUDA], [EDAD MORA],
[NIVEL_RIESGO])

OPENROWSET ('[Provider='NombreServidor','u
suario','password', SELECT ([ID], [PROFESION],
[DEUDA], [EDAD MORA], [NIVEL_RIESGO])
FROM [CLIENTES])
```

Para este caso específico, la forma de alimentar el DMM es a través de la cláusula INSERT–SELECT, donde en primera instancia se realiza la obtención de los datos a través de una consulta SQL por OLE SQL desde una vista minable.

Una vez el modelo posea la información que hemos adecuado al mismo, es hora de usarlo para analizar los datos. La consulta al modelo se hará similar al de una TABLA común de BD.

```
SELECT [ID], [DEUDA], [EDAD MORA], Ries-
goDeCredito. NIVEL_RIESGO

PredictProbability(RiesgoDeCredito. NIVEL_
RIESGO)
```

FROM RiesgoDeCredito PREDICTION JOIN CLIENTES

ON RiesgoDeCredit.Profession=CLIENTES. Profession

AND RiesgoDeCredit. DEUDA = CLIENTES. DEUDA

AND RiesgoDeCredit. EDAD MORA = CLIENTES. EDAD MORA

Esta consulta final al modelo DMM es la que nos permite el análisis de los datos para usarlos en nuestras predicciones futuras y tomar estas como herramienta potencial en la toma de decisiones.

5.2 Lógica difusa

Cuando las decisiones por tomar se encuentran en situaciones donde los criterios de decisión no son evidentes y la incertidumbre es mayor a la deseable se incorpora el concepto de *lógica difusa* para esclarecer el análisis de los datos seguido de los modelos que la lógica borrosa incorpora.

The Fuzzy Logic se agrega como una de las herramientas más utilizadas, tanto en el campo de la inteligencia de negocios campos, como en el de reconocimiento de patrones, Data Mining, estadística aplicada, segmentación de clientes y otros campos, lo que en últimas constituye la aplicación de algoritmos de agrupamiento que permiten identificar las clases que residen en los datos[4], y esclarecer "el proceso no trivial de identificar patrones en datos que sean válidos, novedosos, potencialmente útiles y, por último, comprensibles" [12].

5.2.1. Modelo matemático

Parte de la teoría intuitiva de conjuntos en lógica: sea U el conjunto, posiblemente infinito, de todas las proposiciones. Sean p, q, r, s… sus elementos; es decir, proposiciones atómicas.

En lógica clásica presuponemos una aplicación v del conjunto U en el conjunto {0,1}, de tal manera que v(p)=0 cuando p es falsa y v(p)=1 cuando p es verdadera.

Dado esto, se establece una clasificación de las proposiciones mediante la relación de equivalencia, como sigue:

$$p = q \qquad si \qquad v(p) = v(q)$$

En la teoría axiomática de la probabilidad definimos los espacios de probabilidad por tres factores: conjunto no vacío de los resultados (□), conjunto de sucesos o eventos como parte de W (a) y una función (p) en el intervalo [0,1], que verifica:

a) $0 =< p(A) =< 1$ para todo $A \in a$
b) $p(\Omega) = 1$ y $p(\Phi) = 0$
c) $p(A) =< p(B)$ si $A \subset B$
d) $p(A \cup B) = p(A) + p(B)$ si $A \cap B = \Phi$

Así definida la probabilidad también presupone una estructura de álgebra de Boole de conjuntos. No obstante, la determinación de la probabilidad (la función p) deberá desarrollarse mediante métodos apropiados [7], aunque los casos de aleatoriedad, serán los que menos interés tengan desde el punto de vista lógico.

Para los Conjuntos borrosos y predicados vagos, sea el conjunto P de los predicados: {P, Q, R…} y sea el conjunto U de los objetos del universo del discurso: {x, y, z…}. El producto cartesiano P x U representa el conjunto de todas las proposiciones diádicas {Px, Py, Pz…Qx, Qy, Qz… Rx, Ry, Rz…}. Sólo podemos asignar valores de verdad en el conjunto producto, ya que tanto predicados como objetos carecen de valores veritativos. Mediante el criterio semántico de verdad podemos establecer la clase de los predicados verdaderos

[4] http://cabierta.uchile.cl/revista/31/mantenedor/sub/revisiones_2.pdf

(V) y falsos (F) para cualquier predicado. Por ejemplo, B:

$$V = \{Bx;\ x \text{ es } B\} \qquad F = \{Bx;\ x \text{ no es } B\}$$

Pero sucede que algunas proposiciones no son claramente verdaderas o falsas, sino que tienen una cierta graduación. Deberemos definir, pues, más clases que las F y V clásicas; debemos definir el grado de compatibilidad de B con x. Para ello estableceremos el subconjunto borroso P, según la siguiente expresión:

$$x \in a\ P \qquad \text{si y sólo si} \qquad mp(x) = a$$

En la que el símbolo $\in a$ significa 'pertenece en grado a' y la función mp representa el valor veritativo de la proposición Px para todo $x \in U$, y cuyo campo de existencia es [0,1]. Pero, ¿cómo es la función de pertenencia mp? Ahí radica el problema fundamental: los criterios por seguir para establecer el grado de asignación.

a. Asignación de grados de pertenencia.

b. Proceso individual o colectivo de asignación directa.

c. Procesos probabilísticos y estadísticos.

d. Procesos de análisis de alternativas.

e. Procesos de medición directa o indirecta.

f. Proceso de Zhang.

También es algo problemática la definición de los operadores en el conjunto. En [15] se define la operación conjunción "A y B" o simbólicamente A o B mediante la fórmula:

$$(A \cap B)(x) = \text{Min}\,[A(x), B(x)]$$

Esta fórmula cumple con las propiedades asociativa, conmutativa, idempotente, creciente, continua con 1 como unidad y 0 como absorbente, pero que no cumple otras como A o $(1 \cap A) = Æ$. Se han propuesto otras fórmulas para la definición de la fórmula de la conjunción basadas en la distribución conjunta de probabilidades de A y B que cumplen las propiedades razonables de dicho operador. La negación se define mediante una biyección involutiva decreciente, de tal manera que N: [0,1] ® [1,0], y a partir de esta se puede obtener la regla de la disyunción.

Toda la construcción nos aleja de la estructura de álgebra de Boole, que como hemos visto garantizaba la conversión de las reglas lógicas en operaciones de álgebra de conjuntos. Por ello será conveniente establecer criterios de medición de la borrosidad (entropía borrosa) y sistemas de conversión de los subconjuntos borrosos en sus subconjuntos clásicos más próximos [16].

Dado X como la matriz de los valores de los objetos conocidos, donde cada fila contiene los valores de los objetos $\{x_1^t, \ldots, x_{n_1}^t\}$ y dado Y como la matriz de objetos agregados, donde cada fila contiene los valores de los objetos agregados al conjunto de datos $\{y_1^t, \ldots, y_{n_2}^t\}$, se calcula la distancia entre los centros de las clases *i* y *j* de los objetos conocidos y, por lo tanto, son los centros generados con X.

$$d(v_i, v_j)\ \ \forall i \neq j, \quad i,j \in \{1, \ldots, c\}$$

La distancia entre el objeto agregado *k y* y el centro *i v* de los objetos conocidos, la denotamos

$$d_{ik} = d(y_k, v_i)$$

Se define la función indicatriz:

$$1_{NPC}(y_i) = \begin{Bmatrix} 1 & y_i \ \textit{no presenta cambio} \\ 0 & y_i \ \textit{presenta cambio} \end{Bmatrix} \quad (1)$$

para representar si los datos presentan cambios observados. También denotamos la indicatriz

$$1_{Ci}(y_i) = \begin{Bmatrix} 1 \;\; y_j \;\; es\ asignada\ a\ la\ clase\ i \\ 0 \;\; y_j \;\; si\ no\ es\ asignada\ a\ la\ clase\ i \end{Bmatrix} \quad (2)$$

para la definición de los criterios para la detección de los cambios en los datos que se presentan en las siguientes condiciones:

Condición 1.

$$\alpha_1 \leq \mu_{ik} \leq \alpha_2 \quad \forall k \in \{1,..,n_2\} \quad \forall i \in \{1,..,c\}$$

Condición 2.

$$d_{ik} > \left(\frac{1}{2}\right)\min\{d(v_i,v_j)\} \;\; \forall k \in \{1,..n_2\} \;\; \forall i \neq j \in \{1,..,c\}$$

Seguidas de las etapas de reconocimiento de los estados de los posibles cambios asociados a los datos:

Condición 1.

$$\frac{\sum_{i-1}^{n_2} 1_{NPC}(y_i)}{n_2} > \beta \;\; , \;\; con \;\; 0 < \beta < 1$$

Condición 2 (total datos conocidos y agregados).

$$\frac{\sum_{i-1}^{n_2} 1_{NPC}(y_i)}{n_1 + n_2} > \beta \;\; , \;\; con \;\; 0 < \beta < 1$$

Por último, la etapa final está dada por las decisiones sobre las opciones a consideración frente a las incertidumbres concebidas al inicio del planteamiento del problema y el grado de certidumbre que arrojarán las conclusiones finales:

$$v_i^* = \frac{\sum_{K-1}^{n_2} 1_{NPC}(y_k)(\mu_{ik})^m y_k}{\sum_{K-1}^{n_2} 1_{NPC}(y_k)(\mu_{ik})^m} \;\; 1 \leq i \leq c \qquad (3)$$

Donde V_i^* indica el grado de certidumbre sobre la base del conocimiento, y V_i es el complemento, es decir la incertidumbre que arrojan los datos.

$$v_i = (1 - \lambda_i)v_i + \lambda_i v_i^* \qquad (4)$$

Por último, como elemento final Vi nos determina el punto intermedio en el cual se encuentra nuestra información, ayudándonos a tomar decisiones mucho más cercanas a la realidad y alejadas de los parámetros lineales que no existen en el mundo real; por el contrario, nos muestra un camino que incluye los grados de certidumbre e incertidumbre, y nos ayuda en una toma de decisiones bastante acoplada a la realidad, certera y realista.

$$\lambda_i = \frac{\sum_{j-1}^{n_2} 1_{Ci}(y_j)}{n_1 + n_2} \qquad (5)$$

$$\lambda_i = \frac{\sum_{j-1}^{n_2} 1_{Ci}(y_j)(\mu_{ij})}{n_1 + n_2} \qquad (6)$$

$$\lambda_i = \frac{\sum_{j-1}^{n_2} (1_{Ci}(y_j)\, 1_{NPC}(y_j)(\mu_{ij}))}{\sum_{j-1}^{n_1} (1_{Ci}(x_j)(\mu_{ij})) + \sum_{j-1}^{n_2} (1_{Ci}(y_j)\, 1_{NPC}(y_j)(\mu_{ij}))}$$

6. Conclusiones

Las bodegas de datos son una herramienta necesaria y muy ventajosa para las empresas con respecto a la toma de decisiones; además, representan un

instrumento para ayudar a optimizar el costo/beneficio y obtener la mayor productividad no sólo en términos económicos, sino financieros, humanos, culturales y en general todos los que abarquen el proceso empresarial.

La inteligencia de negocios que comprenden intrínsecamente los Data Warehouses establece como prioridad satisfacer las necesidades de los requerimientos internos de la empresa y además de ello proporcionar información de valioso nivel, no solo a los altos ejecutivos, sino también a todos los empleados que necesiten de este recurso.

Con la implementación de la tecnología Data Warehouse sustentada en Data Marts se planea marcar la diferencia e imponer las pautas a la hora de la toma de decisiones, y además posicionar sus procesos basados en una herramienta robusta que pueda ser usada en tiempo real y que permita al Data Warehouse minimizar el tiempo de análisis de la información y maximizar la velocidad de respuesta a un determinado suceso.

Una de las aplicabilidades de las bodegas de datos es la de satisfacer los requerimientos de información internos de una empresa y gestionarlos de la mejor manera, tanto efectiva como eficiente, para obtener mayor competitividad a la hora de la toma de decisiones.

Data Warehouse es una herramienta sólida y capaz de proveer soluciones a los problemas y requerimientos por parte de la organización, de manera que se convierta en un concepto intrínsecamente asociado a la toma de decisiones.

La utilización de la metodología CRISP en la construcción de un Data Warehouse proporciona confiabilidad, robustez y estandarización de los procesos, de tal manera que facilita la creación de futuros proyectos que posean características similares, ya que genera plantillas uniformes para el proyecto en cuestión. Además de lo anterior, facilita la planificación y dirección del proyecto, de tal manera que se logre hacer un correcto y buen seguimiento del mismo.

La lógica difusa se aplica en la inteligencia de negocios para esclarecer el análisis de los datos, seguido de los modelos que esta misma incorpora cuando las decisiones por tomar se encuentran en situaciones donde los criterios de decisión no son evidentes y la incertidumbre es mayor a la deseable.

Referencias bibliográficas

[1] J. Bustos, "Business Intelligence y Data Warehousing en Window", Danysoft. [En línea]. Disponible: http://www.danysoft.com/free/BIyDW.pdf

[2] Oracle, "Database Machine Warehouse Architectural Comparisons". [En línea]. Disponible: http://www.dmreview.com/

[3] D. Chrysler, then D. Benz, Cross Industry Standard Process. [En línea]. Disponible: http://www.crisp-dm.org/Process/index.htm

[4] BI-SPAIN, "Portal en español sobre software para Business". [En línea]. Disponible: http://www.bi-spain.com

[5] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer y R. Wirth. CRISP-DM 1.0 Step-by-step Data Minig Guide. [En linea]. Disponible: http://www.crisp-dm.org/Process/index.htm, 2000.

[6] J. Galindo, "Conjuntos y Sistemas Difusos". [En línea]. Disponible: http://www.lcc.uma.es/~ppgg/FSS/FSS8.pdf

[7] J. Gondarnores, "Metodologías para Proyectos de Data-Mining. [En línea]. Disponible: http://www.estadistico.com/arts.html?20040426

[8] J. Hernández, "Minería de Datos". [En línea]. Disponible: http://www.estadistico.com/arts.html?20040426

[9] J. Hernández, "Otros aspectos de la Minería de Datos". [En línea]. Disponible: http://www.dsic.upv.es/~jorallo/master/dm5.pdf

[10] C. Wolff, "La Tecnología Data Ware Housing". [En línea]. Disponible: http://www.utpl.edu.ec/eva/descargas/material/140/INFAII21/G4181003.pdf

[11] J. Argos), "Almacenamientos del Data Warehouse (1)". [En línea]. Disponible: http://todobi.blogspot.com/2005_05_01_todobi_archive.html

[12] M. Fayyad, "Data Mining and Knowledge Discovery: Making Sense Out of Data", *IEEE Expert, Intelligent Systems & Their Applications*, vol. 11, no. 4, pp. 20-25, Oct. 1996.

[13] J. González, "Estrategias de Inteligencia de Negocios". [En línea]. Disponible: http://www.cainco.org.bo/es/ecainco/actividades/seminarios/sem18042005/TallerBI.pdf#search=%22data%20mart%2Bdefinicion%2Bralph%20kimball%22

[14] F. Creso, "Agrupamiento Dinámico con Lógica Difusa". [En línea]. Disponible:http://cabierta.uchile.cl/revista/31/mantenedor/sub/revisiones_2.pdf

[15] B. Kosko, E. Trillas "Lógica Difusa". [En línea]. Disponible: http://personal.telefonica.terra.es/web/mir/ferran/kosko.htm

[16] FROILANSOLISALMONACID, "Datawarehouse". [En línea], Disponible: http://www.monografias.com/trabajos6/dawa/dawa.shtml