



Modelamiento de Base de Datos para la Gestión

Descripción y Análisis del modelamiento de las bases de datos para gestión, esto abarcará los repositorios como lo son los Datamart y Datawarehouse y las nuevas tecnologías transaccionales como OLAP, OLTP incluyendo a la Datamining y su análisis para la gestión de una organización.

RONALD IVÁN VACA POQUIOMA
www.seccperu.org/ivanvaca

Modelamiento de Base de Datos para la Gestión

Ronald Iván Vaca Poquioma^{1,2}

¹Sociedad de Estudiantes de Ciencia de la Computación

²Universidad Nacional de Trujillo

ivan@seccperu.org

Resumen

En el presente paper hablaremos acerca del modelamiento de las bases de datos para gestión, esto abarcará los repositorios como lo son los Datamart y Datawarehouse y las nuevas tecnologías que se están imponiendo para el análisis de los datos estables y transaccionales como lo es sistemas basado en OLAP, OLTP incluyendo a la Datamining como uno de tópicos y aplicaciones más resaltante en cuanto al análisis de bases de datos, con el fin de encontrar soluciones claras para la toma de decisiones en la gestión de una organización.

1. Introducción

Hoy en día en las organizaciones requieren no sólo del almacenamiento de datos sobre sus operaciones. Si no también que está sucediendo con la empresa y la sociedad. Para ello toma como referencia a los datos como los agentes informativos para sus dudas. Por ello hoy en día las bases de datos juegan un rol muy importante para la toma de decisiones y ello conlleva a que se desarrollen y construyen grandes repositorios (Datamarts y Datawarehouse) y sistemas que permitan analizar estos datos con el fin de hacer un estudio del comportamiento del mercado a fin de: Gestionar, predecir, analizar, clasificar comportamientos y datos del cliente. Por ello es importante conocer acerca de que tipo de modelos de base de datos nos permiten obtener buenos resultado y sobretodo nos permitirá un mayor desempeño de cualquier sistema de gestión, análisis o de predicción de datos.

2. Conocimientos Previos

2.1. Gestión:

El concepto de gestión hace referencia a la acción y al efecto de gestionar o de

administrar. Gestionar es realizar diligencias conducentes al logro de un negocio o de un deseo cualquiera. Administrar, por otra parte, consiste en gobernar, dirigir, ordenar, disponer u organizar. El término gestión, por lo tanto, implica al conjunto de trámites que se llevan a cabo para resolver un asunto o concretar un proyecto. La gestión es también la dirección o administración de una empresa o de un negocio. [5]

2.2 Base de Datos:

Una base de datos o banco de datos es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso. En este sentido, una biblioteca puede considerarse una base de datos compuesta en su mayoría por documentos y textos impresos en papel e indexados para su consulta. [1]

2.3 Sistema de Soporte de Decisión

Es un conjunto de programas y herramientas que permiten obtener de manera oportuna la información que se requiere mediante el proceso de la toma de decisiones que se desarrolla en un ambiente de incertidumbre. Ayudan a la toma de decisiones de los administradores al combinar datos, modelos analíticos sofisticados y software amigable en un solo sistema poderoso que puede dar soporte a la toma de decisiones semi-estructuradas o no estructuradas.

2.4 Repositorios para Gestión:

2.4.1 Datamart:

Un Datamart es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento.

Un Datamart puede ser alimentado desde los datos de un Datawarehouse, o integrar por si mismo un compendio de distintas fuentes de información. [2]

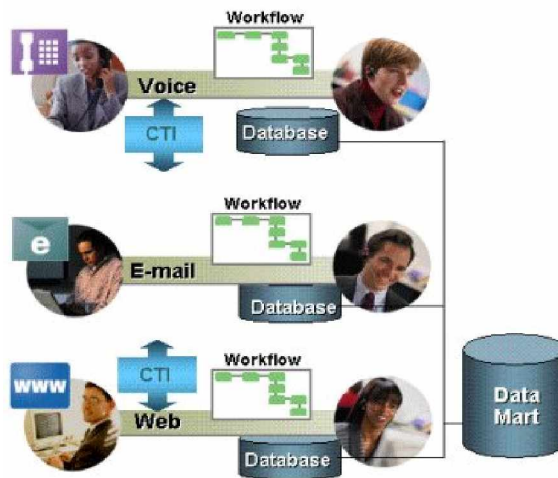


Fig. 1: Esquema Figurativo de un Datamart.

2.4.2 Datawarehouse

Un Datawarehouse es una base de datos corporativa caracterizada por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con gran velocidad de respuesta.

La creación de un Datawarehouse representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de Business Intelligence.

Un Datamart almacena la información de un área o departamento específico y un conjunto de Datamart forman un Datawarehouse. [3]

La ventaja principal de este tipo de bases de datos radica en las estructuras en las que se almacena la información (modelos de tablas en estrella, en copo de nieve, cubos relacionales, etc.).

Este tipo de persistencia de la información es homogénea y fiable, y permite la consulta y el tratamiento jerarquizado de la misma (siempre en un entorno diferente a los sistemas operacionales). [2]

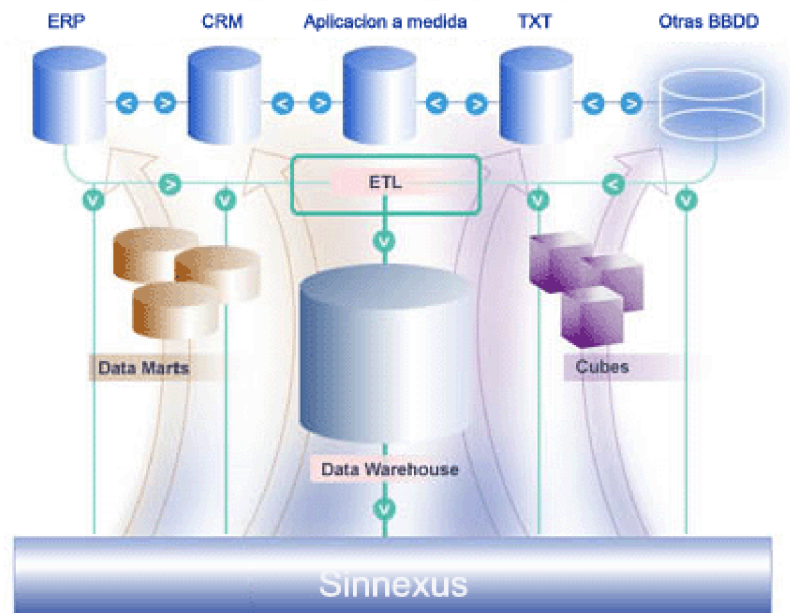


Fig. 2: Gráfico Figurativo de un Datawarehouse

2.5 Bases de datos para Gestión

2.5.1 Base de Datos OLAP (On-Line Analytical Processing)

Los sistemas OLAP son sistemas de bases de datos orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos... etc. Este sistema es típico de los Datamarts. [2]

Los cubos, las dimensiones y las jerarquías son la esencia de la navegación multidimensional del OLAP. Al describir y representar la información en esta forma, los usuarios pueden navegar intuitivamente en un conjunto complejo de datos.

Sin embargo, el solo describir el modelo de datos en una forma más intuitiva, hace muy poco para ayudar a entregar la información al usuario más rápidamente.

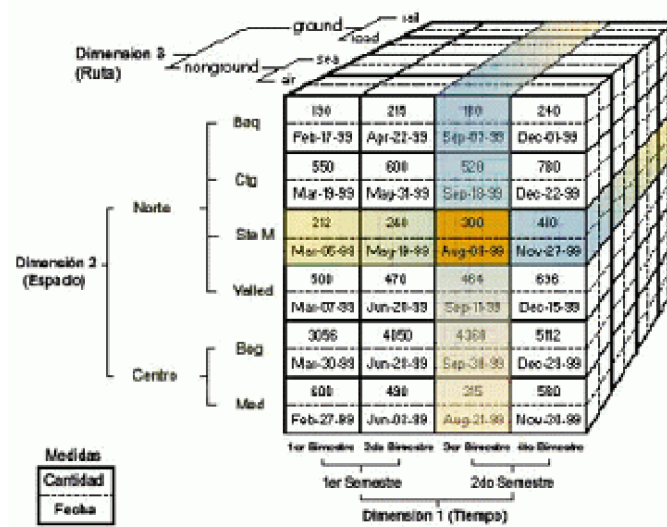


Fig. 3: Cubo dimensional

En los primeros días de la tecnología OLAP, la mayoría de las compañías asumía que la única solución para una aplicación OLAP era un modelo de almacenamiento no relacional. Después, otras compañías descubrieron que a través del uso de estructuras de base de datos (esquemas de estrella y de copo de nieve), índices y el almacenamiento de agregados, se podrían utilizar sistemas de administración de bases de datos relacionales (RDBMS) para el OLAP.

Estos vendedores llamaron a esta tecnología OLAP relacional (ROLAP). Las primeras compañías adoptaron entonces el término OLAP multidimensional (MOLAP), estos conceptos, MOLAP y ROLAP, se explican con más detalle en los siguientes párrafos. Las implementaciones MOLAP normalmente se desempeñan mejor que la tecnología ROLAP, pero tienen problemas de escalabilidad. Por otro lado, las implementaciones ROLAP son más escalables y son frecuentemente atractivas a los clientes debido a que aprovechan las inversiones en tecnologías de bases de datos relacionales preexistentes.

2.5.2 Base de Datos OLTP (On-Line Transactional Processing)

Los sistemas OLTP son sistemas de bases de datos orientadas al procesamiento de transacciones. Una transacción genera un proceso atómico (que debe ser validado con un commit, o invalidado con un rollback), y que puede involucrar operaciones de inserción, modificación y borrado de datos. El proceso transaccional es típico de las bases de datos operacionales. [2]

2.6 Datamining

El Datamining (minería de datos), es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Lo que pretende hacer el Datamining es tratar de ayudar a comprender el contenido de un repositorio de datos. Para ello hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales.

De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación que surge entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento. Vea más diferencias entre datos, información y conocimiento. [2]

3. Modelamiento de Bases de datos para Gestión

3.1 Modelo de Datos de un Datawarehouse

3.1.1 Modelo Relacional

- Maneja la redundancia fuera de los datos. Por lo tanto realizar un cambio en la base significa tocarla en un solo lugar.
- Divide los datos en entidades, las que son representadas como tablas en una base de datos.
- Los MER crecen fácilmente, haciéndose más y más complejos.
- Se puede apreciar la existencia de muchos caminos para ir de una tabla a otra. Sería natural pensar que al tener diversos caminos para llegar desde una tabla a otra, cualquiera de ellos entregaría el mismo resultado, pero lamentablemente esto no siempre sucede así.
- El diagrama se visualiza simétrico, donde todas las tablas se parecen, sin distinguir a priori la importancia de unas respecto a otras. No es fácil de entender tanto para usuarios como para los diseñadores. [4]

3.1.2 Modelo Multidimensional

En general, la estructura básica de un DW para el Modelo Multidimensional está definida por dos elementos: esquemas y tablas.

Tablas de un Datawarehouse:

Como cualquier base de datos relacional, un DW se compone de tablas. Hay dos tipos básicos de tablas en el Modelo Multidimensional:

Tablas Fact (Tabla de Hechos):

Es la tabla central en un esquema dimensional. Es en ella donde se almacenan las mediciones numéricas del negocio. Estas medidas se hacen sobre el grano, o unidad básica de la tabla. El grano o la granularidad de la tabla queda determinada por el nivel de detalle que se almacenará en la tabla. Por ejemplo, para el caso de producto, mercado y tiempo antes visto, el grano puede ser la cantidad de madera vendida 'mensualmente'. El grano revierte las unidades atómicas en el esquema dimensional.

Cada medida es tomada de la intersección de las dimensiones que la definen. Idealmente está compuesta por valores numéricos, continuamente evaluados y aditivos. La razón de estas características es que así se facilita que los miles de registros que involucran una consulta sean comprimidos en unas pocas líneas en un set de respuesta. La clave de la tabla fact recibe el nombre de clave compuesta o concatenada debido a que se forma de la composición (o concatenación) de las llaves primarias de las tablas dimensionales a las que está unida. Así entonces, se distinguen dos tipos de columnas en una tabla fact: columnas fact y columnas key. Donde la columna fact es la que almacena alguna medida de negocio y una columna key forma parte de la clave compuesta de la tabla.

Tablas Lock_up o Dimensionales:

Estas tablas son las que se conectan a la tabla fact, son las que alimentan a la tabla fact. Una tabla lock_up almacena un conjunto de valores que están relacionados a una dimensión particular. Tablas lock_up no contienen hechos, en su lugar los valores en las tablas lock_up son los elementos que determinan la estructura de las dimensiones. Así entonces, en ellas existe el detalle de los valores de la dimensión respectiva. Una tabla lock_up está compuesta de una

primary key que identifica unívocamente una fila en la tabla junto con un conjunto de atributos, y dependiendo del diseño del modelo multidimensional puede existir una foreign key que determina su relación con otra tabla lock_up.

Para decidir si un campo de datos es un atributo o un hecho se analiza la variación de la medida a través del tiempo. Si varía continuamente implicaría tomarlo como un hecho, caso contrario será un atributo. Los atributos dimensionales son un rol determinante en un DDW. Ellos son la fuente de todas las necesidades que debieran cubrirse. Esto significa que la base de datos será tan buena como lo sean los atributos dimensionales, mientras más descriptivos, manejables y de buena calidad, mejor será el DDW.

Esquemas DW

La colección de tablas en el DW se conoce como Esquema. Los esquemas caen dentro de dos categorías básicas: esquemas estrellas y esquemas snowflake. [4]

Esquema Estrella (Star)

El modelo multidimensional también se conoce con el nombre de esquema estrella, pues su estructura base es similar: una tabla central y un conjunto de tablas que la atienden radialmente. (ver Fig. 4).

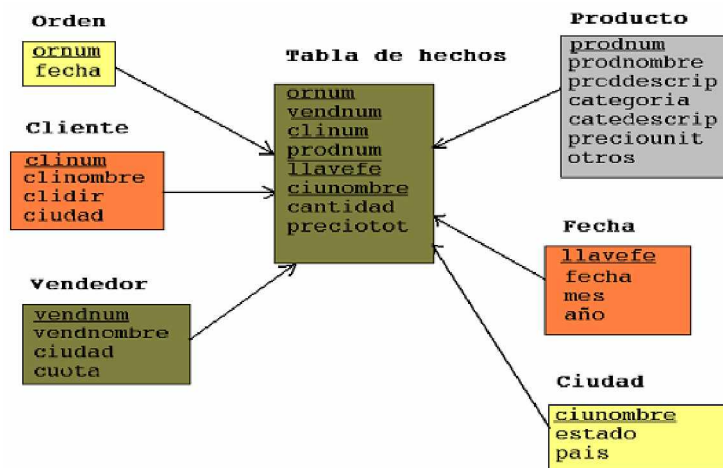


Fig. 4: Gráfica del Esquema Estrella para una base de datos con dimensiones Tiempo, producto, mercado y Cliente.

Esquema Copo de Nieve (Snowflake)

La diferencia del esquema snowflake comparado con el esquema estrella, está en la estructura de las tablas lock_up: las tablas lock_up en el esquema snowflake están normalizadas. Cada tabla lock_up contiene sólo el nivel que es clave primaria en la tabla y la foreign key de su parentesco del nivel más cercano del diagrama. [4]

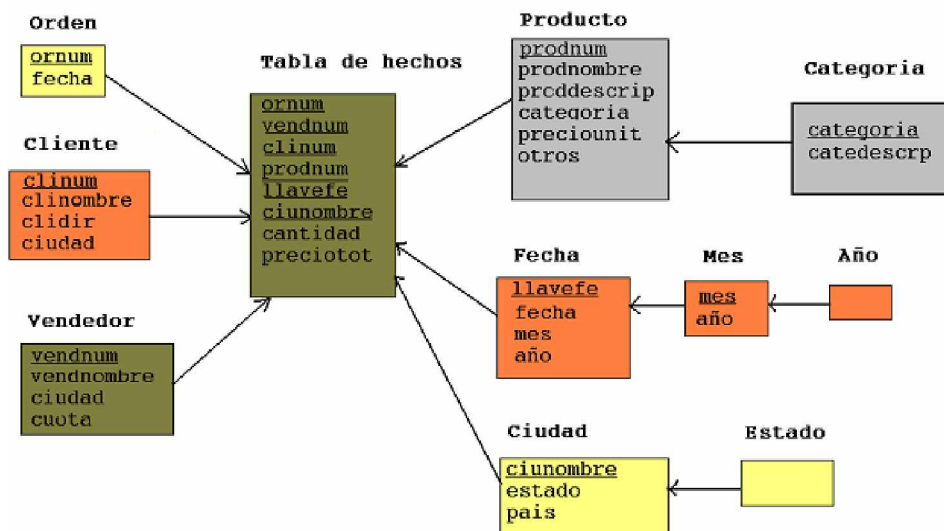


Fig. 5: Gráfica del Esquema Copo de Nieve (Snowflake).

Pasos Básicos del Modelamiento Multidimensional

1. Decidir cuáles serán los procesos de negocios a modelar, basándose en el conocimiento de éstos y de los datos disponibles. Ejemplo: Gastos realizados por cada mercado para cada ítem a nivel mensual. Productos vendidos por cada mercado según el precio en cada mes.
2. Decidir el Grano de la tabla Fact de cada proceso de negocio. Ejemplo: Producto x mercado x tiempo. En este punto se debe tener especial cuidado con la magnitud de la base de datos, con la información que se tiene y con las preguntas que se quiere responder. El grano decidirá las dimensiones del DDW. Cada dimensión debe tener el grano más pequeño que se pueda puesto que las preguntas que se realicen necesitan cortar la base en caminos precisos

(aunque las preguntas no lo pidan explícitamente).

3. Decidir las dimensiones a través del grano. Las dimensiones presentes en la mayoría de los DDW son: tiempo, mercado, producto, cliente. Un grano bien elegido determina la dimensionalidad primaria de la tabla fact. Es posible usualmente agregar dimensiones adicionales al grano básico de la tabla fact, donde estas dimensiones adicionales toman un solo valor para cada combinación de las dimensiones primarias.
4. Si se reconoce que una dimensión adicional deseada viola el grano por causar registros adicionales a los generados, entonces el grano debe ser revisado para acomodar esta dimensión adicional.
5. Elegir las mediciones del negocio para la tabla fact. Se deben establecer los ítems que quedarán determinados por la clave compuesta de la tabla fact. [4]

4. Modelo de Gestión para un Datamining

El proceso del Datamining consta en aplicar a una determinada base de datos las operaciones requeridas de selección, exploración, muestre, transformación y métodos de modelado para extraer patrones y posteriormente evaluarlos para identificar el conjunto de ellos que representarán el conocimiento.

El proceso de KDD (Knowledge Discovery in Databases) es un proceso iterativo porque incluye numerosos pasos en los que el usuario tiene que tomar decisiones. Es iterativo porque puede ser necesario acceder desde una fase a cualquiera de las anteriores, e iterativo porque el proceso es supervisado por el usuario de forma directa.

El proceso consta de cuatro fases:

1. Selección de Objetivos
2. Preparación de Datos
3. Construcción del Modelo
4. Análisis de Resultados

En la fase de construcción del modelo, es la fase central del proceso de descubrimiento en las que se aplican los algoritmos de búsqueda del

conocimiento a los datos previamente preparados.

La definición del modelo a aplicar depende del problema a resolver (meta buscada) y del tipo de datos con los que estamos tratando en cada momento. Según los problemas estos los podemos clasificar en:

- Problemas descriptivos
- Problemas predictivos

4.1 Problemas Descriptivos

Son aquellos problemas cuya meta es simplemente encontrar una descripción de los datos de estudio. Ejemplos:

- Conocer cuáles son los clientes de una organización (características de los mismos).
- Encontrar los productos que frecuentemente se compran juntos.
- Síntomas de enfermedades que se presentan juntas.

Los problemas descriptivos se pueden subdividir en:

- Análisis de Segmentación.
- Análisis de Asociaciones.

4.1.1 Análisis de Segmentación

Consta en encontrar grupos homogéneos en la población de objetos de origen.

4.1.2 Análisis de Asociaciones:

Se persigue obtener relaciones entre los valores de atributos de una base de datos. El ejemplo típico es el de analizar la canasta de compras del cliente.

4.2 Problemas Predictivos (O aprendizaje supervisado en entornos de I.A.)

Son aquellos cuya meta es obtener un modelo que en un futuro pueda ser aplicado para predecir comportamientos.

Los problemas predictivos se pueden subdividir en:

- Problemas de clasificación
- Problemas de predicción de valores

4.2.1 Problemas de Clasificación

Cuando la variable a predecir tiene un número finito de valores

4.2.2 Problemas de Predicción de valores

Cuando la variable a predecir es numérica. Ejemplo: La probabilidad de que un cliente que hace préstamo lo devuelva. Para el modelamiento de estos datos dependiendo del tipo del problema al que se requiera existen técnicas que nos permiten modelar los datos del Datamining, entre los cuales tenemos.

4.3 Modelos para la Solución de los Problemas del Datamining

4.3.1 Modelo RNA (Redes Neuronales Artificiales)

Las redes de neuronas artificiales son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. En inteligencia artificial es frecuente referirse a ellas como redes de neuronas o redes neuronales.

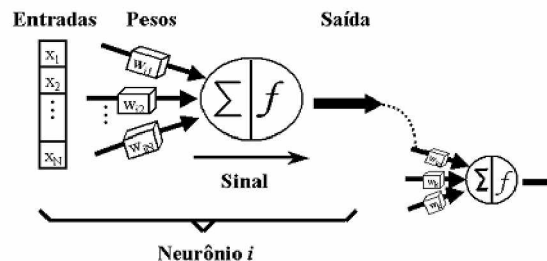


Fig. 6: Modelo de una Red Neuronal Artificial.

Entradas: Estas capas reciben la información desde el exterior como son: Entradas W_j a la neurona.

Pesos: Cada entrada tiene su propio peso relativo el cual proporciona la importancia de la entrada dentro de la función de agregación de la neurona. Ellos son la medida de la fuerza de una conexión de entrada.

Salidas: Cada elemento de procesamiento tiene permitido una única salida $y_i(t)$ que puede estar asociada con un número elevado de otras neuronas. Normalmente, la salida es directamente equivalente al valor resultante de la función de activación.

$$y_i(t) = F_i(a_i(t)) = a_i(t)$$

Algunas topologías de redes neuronales, sin embargo, modifican el valor de la función de transferencia para incorporar un factor de competitividad entre neuronas que sean vecinas. Las neuronas tienen permitidas competir entre ellas, inhibiendo a otras neuronas a menos que tengan una gran fortaleza.

Aquí un ejemplo del modelado de los datos de una RNA en la minería de datos en las finanzas. En éste caso se desea pronosticar algunas variables financieras de una organización y los pasos para modelar los datos son:

1. Identificación de la variable financiera que se va a pronosticar.
2. Construcción y la definición de la base de datos que permitirán activar el proceso de aprendizaje de la Red Neuronal Artificial.
3. Activación del proceso de aprendizaje, con la selección de la arquitectura y los parámetros necesarios para la definición de los pesos de la conexión entre las neuronas.
4. Generalización de los reportes de salida para el pronóstico de la variable financiera.

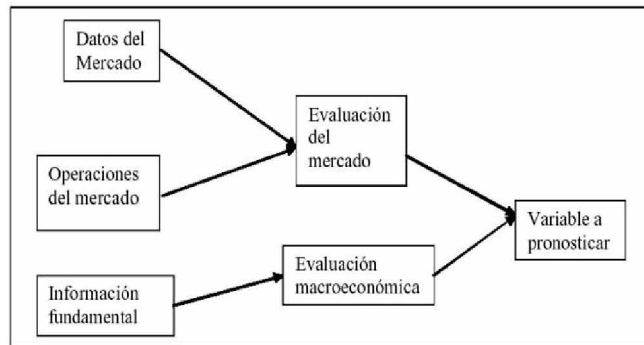


Fig. 7: Modelamiento de un proceso para pronosticar datos de unas variables financieras.

4.3.2 Modelo Algoritmos Genéticos

Son llamados así porque se inspiran en la evolución biológica y su base genético-molecular. Estos algoritmos hacen evolucionar una población de individuos sometiéndola a acciones aleatorias semejantes a las que actúan en la evolución biológica (mutaciones y recombinaciones genéticas), así como también a una Selección de acuerdo con algún criterio, en función del cual se decide cuáles son los individuos más adaptados, que sobreviven, y cuáles los menos aptos, que son descartados.

Funcionamiento de un algoritmos Genético Básico

En general el pseudocódigo consiste en los siguientes pasos:

Inicialización: Se genera aleatoriamente la población inicial, que está constituida por un conjunto de cromosomas los cuales representan las posibles soluciones del problema.

Evaluación: A cada uno de los cromosomas de esta población se aplicará la función de aptitud para saber qué tan "buena" es la solución que se está codificando.

Condición de término: El AG se deberá detener cuando se alcance la solución óptima, pero ésta generalmente se desconoce, por lo que se deben utilizar otros criterios de detención. Mientras no se cumpla la condición de término se hace lo siguiente:

- Selección: Después de saber la aptitud de cada cromosoma se procede a elegir los cromosomas que serán cruzados en la siguiente generación. Los cromosomas con mejor aptitud tienen mayor probabilidad de ser seleccionados.
- Cruzamiento: El cruzamiento es el principal operador genético, representa la re-producción sexual, opera sobre dos cromosomas a la vez para generar dos descendientes donde se combinan las características de ambos cromosomas padres.
- Mutación: modifica al azar parte del cromosoma de los individuos, y permite alcanzar zonas del espacio de búsqueda que no estaban cubiertas por los individuos de la población actual.

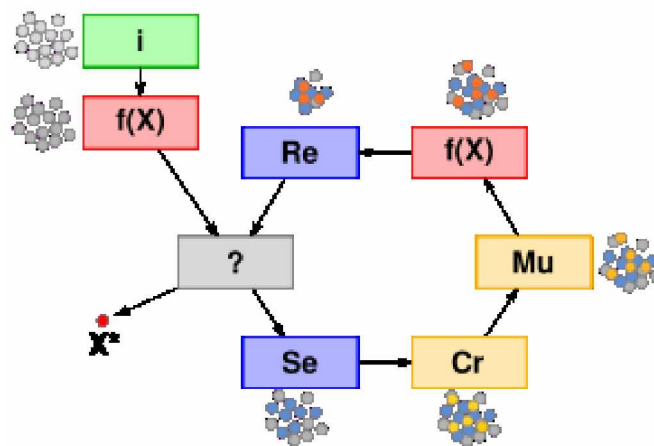


Fig. 6: Algoritmo genético i: inicialización, $f(X)$: evaluación, ? : condición de término, Se: selección, Cr: cruzamiento, Mu: mutación, Re: reemplazo, X^* : mejor solución.

6. Conclusiones

- Podemos concluir que los repositorios de grandes masas de datos como lo son los datamarts y los datawarehouses permiten la gestión de los datos para la adecuada gestión de decisiones de una organización.
- Los sistemas de soporte de decisión tienen como herramientas de aplicación a sistemas basados en Inteligencia Artificial (Redes Neuronales, Algoritmos Genéticos, etc.) y algoritmos probabilísticos.

- El esquema copo de nieve permite mayor robustez en el modelamiento multi-dimensional debido a la normalización.
- El diseño del modelo de un Datawarehouse deberá necesariamente estar definido en forma menos precisa que el diseño de sistemas operacionales.

7. Referencias

- [1] http://es.wikipedia.org/wiki/Base_de_datos
- [2] <http://www.sinnexus.com/>
- [3] Ryner Huamantumba, "*Datamart paso a paso*"
- [4] Carmen Gloria Wolf, "*Modelo Multidimensional*", Agosto del 2002
- [5] <http://www.definicion.de>
- [6] http://es.wikipedia.org/wiki/Red_neuronal_artificial