# APLICACION DE LA MINERIA DE DATOS SOBRE BASES DE DATOS TRANSACCIONALES

## *APLICATION OF DATA MINING ON TRANSACTIONAL DATA BASES*

**Alvaro Troche Clavijo**

**Instituto de Investigaciones de Ciencia y Tecnología**
**Universidad La Salle - Bolivia**

alvarotrochec@hotmail.com

**RESUMEN**

El presente artículo presenta un breve análisis sobre la aplicación de técnicas de minería de datos pero aplicadas sobre ambientes distintos a los repositorios correspondientes a bases de datos analíticas, es decir, cuál sería el impacto de realizar dichos análisis sobre bases de datos transaccionales.

**ABSTRACT**

This article presents a brief discussion on the application of data mining techniques applied to different environments but the repositories corresponding to analytical databases, ie, what the impact of such analyzes on transactional databases. PALABRAS CLAVE: Minería de Datos, KDD, Integridad de información, Data Warehousing, OLAP, OLTP

**Minería de datos.-** Campo de las ciencias orientadas a la informática referido al proceso que intenta descubrir conocimiento a través de patrones en grandes volúmenes de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, que involucra aspectos de bases de datos y gestión de datos, procesamiento de datos, el modelo y las consideraciones de inferencia, métricas de Intereses, consideraciones de la Teoría de la complejidad computacional, post-procesamiento de las estructuras descubiertas, la visualización y actualización en línea.

**KDD.-** (Knowledge Discovery from Databases) considerado un proceso para identificar comportamientos y/o patrones válidos, novedosos, potencialmente útiles

y con la característica de que son comprensibles a partir de los datos. Se pretende también, dicho en otras palabras encontrar conocimiento útil, válido, relevante y nuevo sobre una determinada actividad mediante algoritmos, dadas las crecientes órdenes de magnitud en los datos Integridad de información.- Concepto que hace referencia a la exactitud, y legitimidad que posee un dato o grupos de datos para que los mismos puedan ser usados para otros fines dando como resultado un alto grado de confiabilidad de resultados. Data Warehousing.- Proceso de extraer y filtrar datos de las operaciones comunes de la organización, procedentes de los distintos sistemas de información operacionales, transaccionales y/o sistemas externos, para transformarlos, integrarlos y almacenarlos en un depósito o almacén de datos (Data Warehouse, en inglés) con el fin de acceder a ellos y que sirvan de apoyo al proceso de toma de decisiones de una organización OLAP.- (On-Line Analytical Processing). Es una solución utilizada en el campo de la llamada inteligencia empresarial (o Business Intelligence) cuyo objetivo es agilizar la consulta de grandes cantidades de datos. Para ello utiliza estructuras multidimensionales (o cubos OLAP) que contienen datos resumidos de grandes bases de datos o Sistemas Transaccionales. Se usa en informes de negocios de ventas, marketing, informes de dirección, minería de datos y áreas similares.

**OLTP.-** Es la sigla en inglés de Procesamiento de Transacciones En Línea (OnLine Transaction Processing) es un tipo de sistemas que facilitan y administran aplicaciones transaccionales, usualmente para entrada de datos y recuperación y procesamiento de transacciones (gestor transaccional).

**KEYWORDS:** Data Mining, KDD, data integrity, Data Warehousing, OLAP, OLTP

## 1. INTRODUCCION

Hoy en día se ha popularizado, por así decirlo, la aplicación practica de técnicas de BI y Data Warehouse en muchas instituciones de nuestro medio, tanto en el sector público y tal vez con mayor intensidad y utilidad en el sector privado. Todo este proceso de Data Warehousing según muchas teorías planteadas, sirve de base para realizar procesos de Minería de Datos.

Si bien la minería de datos consiste en la extracción de patrones y modelos con un alto grado de utilidad sobre bases de datos de gran tamaño también requiere que dichas bases cuenten con ciertas características, como las de tener un muy alto grado de consistencia de información (ideal un 100% de nivel de consistencia), y el nivel de normalización del repositorio de datos tiene que ser adecuado. Por lo tanto todo indica que la minería de datos tiene que ser

explotada de bases de datos OLAP, es decir bases de datos pre procesadas con información asociada a temas específicos y con un alto grado de consistencia. De todo lo expuesto anteriormente, podríamos inferir que realizar este proceso de análisis de información sobre bases de datos transaccionales OLTP no tendría sentido debido al nivel de normalización que se maneja, al nivel de inconsistencias que podría llegar a tener mismo que podría generar información no real, y por último la diversidad de información que se maneja ya que una base de datos transaccional a diferencia de una base de datos del tipo analítica, no está orientada a un solo tema en específico, sino a todos los ámbitos que requiere una determinada institución, por lo tanto es información muy variada.

**2. OBJETIVO**

El objetivo del presente artículo es identificar y discutir aspectos que sería importante tomar en cuenta en el caso de que se tenga la necesidad de aplicar técnicas de minería de datos sobre bases de datos que no hayan sido procesadas previamente (bases de datos transaccionales) debido al tiempo que tomaría crear bases de datos alternas con información pre tratada para su análisis OLAP.

Además se pretende resaltar las consecuencias de no aplicar ciertas consideraciones sobre la información antes de que la misma sea utilizada como insumo para inferir resultados.

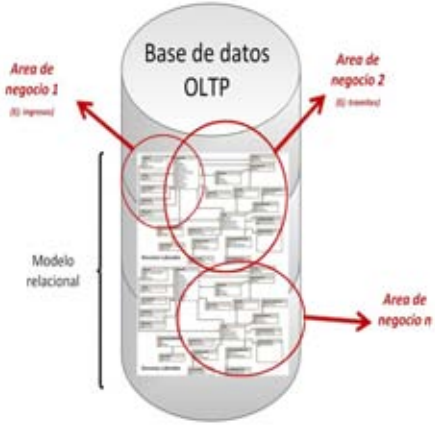
**3. CONTENIDO**

Empecemos primero resaltando la necesidad de almacenar información, independientemente de la naturaleza o negocio a la cual este asociada. La experiencia nos ha enseñado esta información con el tiempo llega a ser un insumo importante para mejora o no cometer los mismos errores de tiempos pasados, es de este aspecto que nace la necesidad de analizar información histórica.

Este análisis debe ser realizado de forma ordenada y sistemática, detallando en primera instancia ¿Qué es lo que quiero analizar?. No olvidemos que cualquiera que sea la naturaleza de la información contenida en una base de datos, siempre esta dividida o puede ser clasificada en "sectores de información". A que nos referimos cuando hablamos de Sectores de información, pues bien, a que no toda la información tiene el mismo fin y puede ser diferenciada por Áreas de Negocio los cuales puedo analizar independientemente.

Pongamos el ejemplo de una universidad, si deseo analizar la información de la misma, lo primero que tengo que hacer es dividir la información en Áreas de negocio como por ejemplo: (1) información administrativa, (2) Información Académica, (3) Información

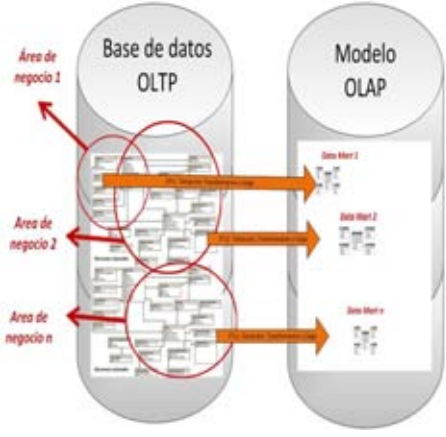
sobre la administración de Aulas, (4) Información sobre los Docentes, etc….(Grafico I)




I.- Clasificación y definición de Áreas de Negocio o Sectores de información. (Fuente Propia)

Después de haber realizado esta clasificación se procede a seleccionar que datos necesito de cada una de esas denominadas "áreas de negocio", con esto nos referimos a: como la quiero agrupar?, que información cuantitativa voy a necesitar?, que información cualitativa requiero?. Todo este proceso nos lleva pues a la construcción de lo que actualmente se conoce como Data Warehouse.

Con este análisis ahora puedo asociar cada una de mis "sectores de información" o "áreas de negocio" con los llamados "Data Marts" que nos son más que subconjuntos de mi DWH con información de cada área pero ya con un pre tratado previo.(Grafico II)




II.- Áreas de negocio asociados a Data Mart. (Fuente Propia)

Es importante recalcar que en el proceso de Extracción, Transformación y Carga de información al DWH, se realizan actividades de

a) Depuración de información inconsistente
b)Identificación de información inconsistente

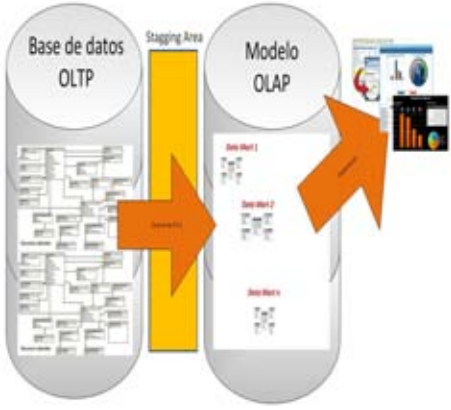
Estos procesos garantizan entonces que cualquier tipo de análisis que se vaya a realizar sobre cualquiera de las áreas de negocio (Data marts) tendrá un alto grado de confiabilidad.

Ya teniendo este repositorio de información pre procesado y con un grado de consistencia muy alto, puedo aplicar distintas técnicas de análisis de información, es entonces donde entra la minería de datos.

Aplicando técnicas de minería de datos puedo entender la información que ahora tengo, identificando ciertos comportamientos que antes eran imperceptibles como la relación entre variables la dependencia entre variables, etc. Todo esto con el fin de "entender" mi información, y de esta forma poder predecir acontecimientos futuros y justificar el comportamiento actual de la información.

Como se observa en el análisis realizado, la minería de datos tiene por detrás otros procesos que ayudan a que la aplicación de estas técnicas tengan un resultado confiable, por lo tanto, este proceso de apoyo enmarquémoslo como Data Warehousing.

El proceso de Data Warehousing, requiere un tiempo considerable para su creación, tomando en cuenta que las etapas para su construcción son similares a un proceso de desarrollo de software, nos referimos a las tareas de Análisis, Diseño, Pruebas, Puesta en producción, citando solo los procesos más básicos.(Grafico III)




III.- Proceso de construcción de un DWH
(Fuente Propia)

Problema del tiempo….
Que sucede si no contamos con este tiempo?..Es necesario ver otras alternativas que me ayuden a cumplir con mis objetivos, que para nuestro caso de estudio es "analizar un sector de información sin contar con un repositorio OLAP disponible".

Empecemos entonces citando los problemas que se presentan en la información origen o transaccional, como quiera llamársele. Esta información generalmente cuenta con información que tiene las siguientes características:

i) Información no clasificada en sectores de información
ii) Información altamente volátil
iii) Información con un determinado nivel de inconsistencia

iv) Información no sumarizada (muy alto nivel de granularidad)

Nos surge la pregunta entonces como subsanar estos aspectos para que podamos realizar directamente los procesos involucrados en la minería de datos sin tener resultados altamente distorsionados.

Subsanando inconvenientes….

En primera instancia debemos definir claramente la información que deseamos analizar, si bien la información no está claramente identificada como en un DWH, puede ser clasificada sin muchos inconvenientes en una base transaccional, la única diferencia que existiría seria la cantidad de estructuras de datos a tomar en cuente. En resumen, la clasificación en "sectores de información" se complica pero es factible.

Como Segundo punto de análisis, en una base de datos transaccional (OLTP) siendo la información altamente cambiante o volátil es necesario tomar medidas para dejar en primera instancia estática la información. Este proceso puede ser realizado utilizando "bases de datos alternas" con la información estática a una fecha. (Grafico IV)

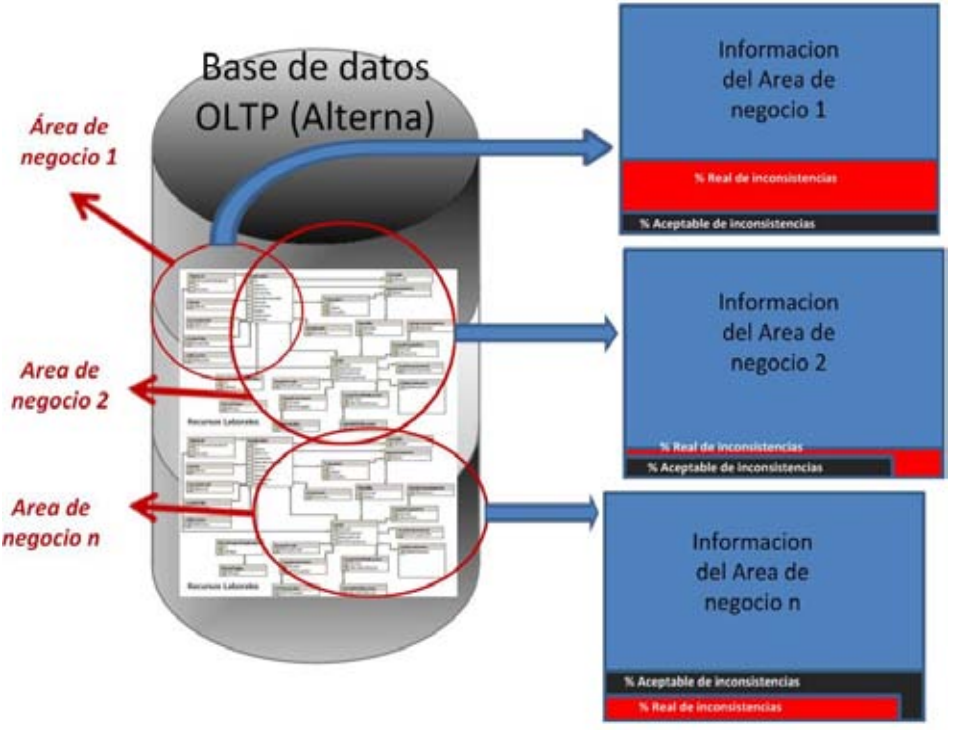
Este proceso contribuiría a solucionar dos problemas: (1) no entorpecer el desenvolvimiento de la base de datos


IV.- Alternativa de acceso a datos con una base de datos alterna (Fuente Propia)

transaccional con procesos de análisis muy complejos que utilizarían muchos recursos del sistema (2) la información no sufriría modificación.

Con relación a la inconsistencia de información, podemos afirmar que este es uno de los problemas más delicados para nuestro estudio, ya que es el principal motivo para que un análisis sea incongruente y muy poco real…

Cómo deberíamos tratar la información inconsistente? Este punto implica un análisis previo sobre qué porcentaje puede considerarse despreciable dependiendo de qué sector de información estemos analizando.(Grafico V)




V.- Porcentaje (%) de inconsistencias y % aceptable de sus existencia
(Fuente Propia)

Si nos referimos a sectores de información que manejan los datos de tipo monetario (ingresos, egresos, recaudaciones, etc…), puede ser que el porcentaje de información inconsistente no pase de un 2.00 %, si nos referimos a universos, por ejemplo de clientes o de transacciones x año, puede ser que ese porcentaje pueda llegar a ser un poco mayor.

Que se logra con esto?, pues bien…. Enfocarnos en depurar la información que necesite y pueda ser depurada para apuntar a estos porcentajes de margen aceptable. Para este proceso es necesario también tomar las siguientes decisiones:

a) Que información inconsistente puedo excluir (por ser considerada información basura)

b) Que información inconsistente puedo depurar (dependiendo si tengo acceso a información origen que me ayude a completar dicha información. Además de velar por la autorización respectiva de los dueños de la información)

c) Que información inconsistente no depurada debo incluir en las estadísticas finales por ser información crítica o de alto grado de importancia.

Tomando en cuenta los aspectos mencionados estoy garantizando que los resultados obtenidos tengan un alto grado de confiabilidad. Pero es importante recalcar que los niveles y porcentajes despreciables mencionados anteriormente, deben ser de conocimiento, respaldados y aceptados por los clientes o en todo caso los usuarios que van a recibir la información para su análisis. Procesos que requieren mucho recurso del procesador….

El ultimo inconveniente que se debería analizar es el nivel de complejidad de las consultas que van a ser ejecutadas. No olvidemos que estamos accediendo a un repositorio de información altamente normalizada, lo que ocasiona que cualquier consulta envíe de 5 a 10 veces más carga de trabajo a un procesador.

Por lo tanto consideremos las siguientes premisas:

i) Se debe optimizar al máximo las consultas desarrolladas, considerando para esto la creación de índices, manejo de particionamiento de estructuras, etc…

ii) Evitar usar sub consultas y otro tipo complejidad en las consultas a datos para acelerar los procesos (Mantener siempre la relación Maestro - Detalle en consultas)

iii) Deshabilitar triggers o disparadores que probablemente hayan estado habilitados en la base de datos de producción, ya que su utilidad en la base de datos alterna ya no es preponderante.

Este tipo de consideraciones no son el recurso único para crear accesos rápidos, pero desde mi humilde punto de vista, son la base para llegar a un acercamiento a las consultas OLAP.

Finalizando podemos aplicar ya sobre esta información pre tratada (entre comillas) procesos de análisis, que desde ningún punto de vista serán exactamente iguales que un análisis de datos realizado directamente sobre repositorios OLAP. Pero el apoyo a la toma de decisiones en mayor o menor grado podrá ser visible. El proceso de "minería de datos" podrá ser realizado en todas sus etapas sobre una base de datos con márgenes de error aceptables, los cuales no deberían afectar en las predicciones y la construcción de modelos predictivos.

## 4. CONCLUSIONES Y RECOMENDACIONES

Para concluir queda mencionar que estos aspectos no son más que parámetros y consideraciones importantes capaces de darme otras posibilidades para aplicar técnicas de "minería de datos" si pasar por ciertas etapas (saltos que muchos expertos considerarían riesgosas), que conllevan a utilizar grandes cantidades de tiempo y recursos, pero que son importantes al fin.

Con el presente análisis no se quiere desmerecer la importancia innegable de la construcción de los repositorios de Data Warehouse ya que como es de conocimiento público, en su proceso está inmerso un aspecto importante que es la de depurar, filtrar y consolidar información en sus procesos denominados ETL, aspecto que es de suma utilidad para cualquier empresa hoy en día.

## 5. BIBLIOGRAFIA

Libros de referencia.-

[1] O'brien James, (2001) "Sistemas de Información Gerencial", Editorial McGraw Hill.

[2] Hernández O. J., Ramírez Q. M. José, Ferri R. C.(2004) "Introducción a la Minería de Datos", Pearson.

[3] Jiawei Han, Micheline Kamber, (2001) "Minería de Datos, Técnicas y conceptos".

[4] Perez Lopez Cesar "Minería de Datos, Técnicas y Herramientas, 2001.

[5] Nevado Cabello Victoria, (2005) "Introducción a las Bases de Datos Relacionales", VisionLibros.

[6] Laudon Keneth C., Laudon Jane P. "Sistemas de Información Gerencial", 8va ed.

[7] Jhonson Joseph, (2002). "Data Base Performance Tunnig", Sibex inc.

[8] http://datawarehouse.ittoolbox.com/

Recibido: 11/07/2013
Aceptado: 09/01/2014