UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

LA MINERÍA DE DATOS COMO EL PROCESO PARA EL DESCUBRIMIENTO Y GENERACIÓN DE CONOCIMIENTO EN UNA BASE DE DATOS

SEMINARIO DE INVESTIGACIÓN INFORMÁTICA

MARÍA ISABEL ANGELES LARRIETA
ANGÉLICA MARÍA SANTILLÁN GÓMEZ

MÉXICO, D.F. 1998

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

LA MINERÍA DE DATOS COMO EL PROCESO PARA EL DESCUBRIMIENTO Y GENERACIÓN DE CONOCIMIENTO EN UNA BASE DE DATOS

SEMINARIO DE INVESTIGACIÓN INFORMÁTICA
QUE PARA OBTENER EL TÍTULO DE:

LICENCIADO EN INFORMÁTICA

PRESENTAN:

MARÍA ISABEL ANGELES LARRIETA
ANGÉLICA MARÍA SANTILLÁN GÓMEZ

ASESOR DEL SEMINARIO:

DOCTOR RICARDO RIVERA SOLER

MÉXICO, D.F. 1998

Agradecimientos

A Dios

Por darme fuerza y valor para alcanzar uno de mis objetivos, mi carrera.

A mis Padres

Que con su infinito amor y apoyo incondicional me han guiado en la vida.

A mis Hermanos

Que creyeron en mi y me ayudaron cuando más lo necesitaba.

A la Universidad Nacional Autónoma de México

Por ser parte de mi formación profesional.

A mi Asesor

Que con su gran experiencia y sabiduría guió mi trabajo de investigación.

A las Marías

Por estar siempre juntas en los momentos más difíciles de la carrera.

María Isabel Angeles Larrieta

Agradecimientos

Señor ayúdame a aprovechar para bien esta oportunidad que hoy me brindas pues será la mejor forma de agradecerte.

A Dios

El ejemplo de lucha constante, el cariño que me han dado, el apoyo que me han brindado, la paciencia que me han tenido y la confianza que en mí han depositado permitieron que lograra concluir con una de mis metas.

A mis padres

Por su gran cooperación, excelente colaboración, apoyo e invaluable amistad.

A mis amigas Isabel y Pilar

Por los momentos de aliento y paciencia que me ha brindado.

Enrique García Portillo

Por ser mi segunda casa y fuente de conocimientos.

A la Universidad Nacional Autónoma de México

Por su excelente orientación y gran ejemplo.

A mi asesor de tesis: Dr. Ricardo Rivera Soler

Angélica María Santillán Gómez.

ÍNDICE

	Página
Agradecimientos	xi
Introducción	xii
1. Marco Problemático	1
1.1 Antecedentes	1
1.2 Identificación del problema	1
1.3 Demarcación del fenómeno	2
1.4 Conocimiento empírico del medio (Observación naturista)	2
1.4.1 Lista de personas a entrevistar	4
1.4.2 Conglomerado y análisis de las respuestas a los cuestionarios aplicados	4
1.4.3 Conclusión general	7
1.5 Opiniones profesionales	8
1.5.1 Lista de personas a entrevistar	8
1.5.2 Conglomerado y análisis de las respuestas a los cuestionarios aplicados	8
1.5.3 Conclusión general	12
1.6 Hipótesis preliminar	12
1.6.1 Presentación del problema en su relación causa-efecto	12
1.6.2 Ejemplos de correlación de hipótesis	14
1.6.3 Hipótesis preliminar	16
1.7Objetivos	16
1.7.1 Personales	16
1.7.2 Particulares	17
1.7.3 Generales	17
2. Marco Teórico	18
2.1 Acopio Bibliográfico	18
2.1.1 Libros	18
2.1.1.1 Diccionarios	19
2.1.2 Tesis	19
2.1.3 Revistas	20
2.1.4 Periódicos	20
2.1.5 Seminarios	21
2.1.6 Conferencias	22
2.1.7 Internet	23
2.1.8 Proveedores	24
2.2 Investigación actualmente desarrollada	25
2.3 Bibliografía en venta	28
2.4 Conclusiones	30

3. Marco Conceptual

Página 31

3.1 Antecedentes 31
3.2 Definiciones 33
3.2.1 Etimológica 33
3.2.2 De diccionario 33
3.2.3 De autores 34
3.2.4 Propia 35
3.3 Sinónimos 35
3.4 Evolución 35
3.4.1 Almacenamiento de los datos (1960s) 36
3.4.1.1 Banco de datos 36
3.4.2 Acceso a los datos 37
3.4.2.1 Base de datos 37
3.4.2.2 DBMS (Data Base Management System) 38
3.4.2.3 Modelos de bases de datos 38
3.4.2.3.1 Sistema Manejador de Archivos (FMS) 39
3.4.2.3.2 Modelo de Base de Datos Jerárquica 39
3.4.2.3.3 Modelo de Base de Datos de Red 40
3.4.2.3.4 Modelo de Base de Datos Relacional 41
3.4.2.4 Arquitecturas de DBMS 41
3.4.2.4.1 Plataforma Centralizada 41
3.4.2.4.2 Sistemas de Bases de Datos Cliente/Servidor 42
3.4.2.4.2.1 La tecnología Cliente/Servidor 43
3.4.2.4.3 Sistemas de Procesamiento Distribuido 44
3.4.2.5 Lenguajes de Programación de Aplicaciones de Bases de Datos 44
3.4.2.5.1 Lenguajes procedurales 44
3.4.2.5.2 Lenguajes de Consulta Estructurados (SQL - Structured Query Language -) 45
3.4.3 Datawarehousing y soporte de decisiones(1990) 46
3.4.4 Minería de datos (hoy en día) 48
3.5 Clasificación 50
3.6 Características generales 50
3.7 Estructura 51
3.7.1 Algoritmos o programas que buscan (mineros) 51
3.7.2 Datos históricos (en dónde buscan) 51
3.7.3 Criterios de búsqueda (qué se busca) 51
3.7.4 Almacenamiento de hallazgos (Cofre de tesoros) 52
3.8 Usuarios de la minería de datos 52
3.9 Areas de aplicación 52
3.10 Proveedores 53
3.11 Plataformas 53
3.11.1 Clasificación 53
3.11.2 Front-Ends 54
3.11.3 Back-Ends 55
3.12 Conclusiones 56

Página

4. Marco Metodológico 57

4.1 Conocer la información de la empresa 57
4.1.1. Conocer la forma de almacenamiento 57
4.1.1.1 Banco de datos 57
4.1.1.2 Base de datos 58
4.1.1.3 Data warehouse 59
4.1.2. Conocer la estructura de almacenamiento de la información 60
4.1.3 Conocer y determinar el volumen de la información 62
4.2 Preparar la información para el proceso de minería de datos 63
4.3. Usar los programas mineros para buscar en la información seleccionada los criterios, ideas, normas o cuestionamientos definidos 64
4.4. Incorporar la información obtenida al proceso de toma de decisiones ... 64
4.5. Medir los resultados 65
4.6. Ejemplos de aplicaciones de extracción de datos 65
4.7 Conclusiones 66

5. Marco Instrumental 67

5.1 Propuesta de acción 67
5.1.1. Publicación de un artículo 67
5.1.2 Diseño de una hoja html de minería de datos 68
5.1.3 Enviar marco metodológico como folleto a universidades 68
5.2 Plan y programa de trabajo 69
5.3 Conclusiones 71

6. Conclusiones generales 72

Anexos

Anexo 1 Diseño del cuestionario de minería de datos 74
Anexo 2 Back Ends 76
Anexo 3 Software para minería de datos 78

Glosario 93

Bibliografía 97

Marco problemático

1.MARCO PROBLEMÁTICO

1.1 Antecedentes

Todas las empresas o la mayoría de ellas, siempre se han preocupado por la obtención de información, veraz y oportuna, que ayude a la gerencia a la toma de decisiones, así como la forma de procesar sus datos de entrada y salida, y el medio en el que han de ser almacenados.

Durante la vida productiva de la mayoría de las empresas, la información que generan se va acumulando sin sentido y con la mínima utilidad, por lo que no es aprovechada y explotada para beneficio propio.

A lo largo de la carrera, Licenciatura en Informática, hemos podido darnos cuenta de la importancia que tiene la información para cualquier tipo de empresa, así como la forma en que esta organizada y almacenada.

Al trabajar con bases de datos, hemos descubierto que se puede explotar la información que se va generando en una organización para multiplicar sus beneficios y productividad, ayudándola a ser más competitiva.

Nuestro objetivo, en un principio, se basó en el diseño de las bases de datos relacionales, ahora nuestro interés va más allá de crear bases de datos y de su forma de almacenar la información, pasando a ser la generación del conocimiento a partir de bases de datos ya existentes, siendo el propósito primordial de la minería de datos.

En nuestro trabajo de investigación se pretende mostrar a la minería de datos como un proceso para el descubrimiento y generación de conocimiento en una base de datos, sus alcances y limitaciones, así como su entorno tecnológico.

La minería de datos como el proceso para el descubrimiento y generación de conocimiento en una base de datos.

1.2 Identificación del problema

Para todas las empresas que manejan grandes volúmenes de información, sin importar su actividad o giro, es muy importante conocer datos adicionales a los generados por sus sistemas, datos que en la mayoría de las veces se quedan almacenados y que no son explotados al máximo. Esta información adicional puede ser obtenida a través de análisis complicados y lentos los cuales requieren de fuertes inversiones de personal, tiempo y dinero lo que representa para la empresa una fuente de pérdida debido al gran desembolso y a la necesidad no satisfecha de información que le ayude a descubrir nuevas oportunidades, haciéndola más competitiva en el mercado.

La falta de conocimiento de la existencia o nula utilización de herramientas que ayuden a analizar, en forma detallada, los datos generados por los sistemas de la empresa ha provocado:

- Carencia de formas de identificación de datos poco comunes o irregulares.
- Grandes períodos invertidos por las personas que realizan los análisis.
- Difícil tratamiento de los criterios para encontrar patrones de conducta debido al crecimiento desmedido de la información en períodos muy cortos.
- No se cuenta con información que ayude a crear estrategias de comportamiento futuro para la empresa.
- La aceptación general de que en las grandes bases de datos existen valores desaprovechados.
- La falta de consolidación de entradas de la base de datos que tiene distintas presentaciones para el usuario.
- La desintegración de las bases de datos.

1.3 Demarcación del fenómeno

La minería de datos es un proceso de propósito general que puede ser aplicado, preferentemente, a empresas que generan grandes volúmenes de información. El enfoque que se le dará durante el desarrollo del trabajo de investigación, será aplicable a todos los sectores empresariales.

1.4 Conocimiento empírico en el medio (observación naturalista)

Con objeto de corroborar el problema identificado y demarcado en nuestro trabajo de investigación recurrimos a un grupo de personas con conocimientos empíricos, personas calificadas no necesariamente profesionistas, para conocer su punto de vista y opinión con respecto a nuestra problemática.

Al efecto diseñamos el siguiente cuestionario (Anexo 1) para ser aplicado, dando a conocer la pregunta, el porque de la pregunta y la respuesta esperada (RE).

1. ¿En la organización en la que labora, utilizan algún modelo que les ayude al diseño de la información?
¿Porqué?: Conocer si hacen uso o no de los modelos de diseño de información.
RE: Si.

2. *Si contestó afirmativamente la pregunta anterior, señale con "X" que modelos usa.*
¿Porqué?: Conocer los modelos del diseño de información utilizados. La minería de datos trabaja, de forma optima, sobre el modelo relacional.
RE: Elección del modelo relacional.

3. ¿En la organización en la que labora, diseñan las bases de datos para su explotación inmediata o futura?
¿Porqué?: Conocer y medir el alcance y aprovechamiento esperado de la información.
RE: Para su explotación inmediata.

4. *¿Usa modelos que le ayuden al análisis de su información histórica?*
¿Porqué?: Saber si se apoyan en modelos de análisis de información para la toma de decisiones.
RE: No.

5. Si contestó afirmativamente la pregunta anterior, señale con "X" qué modelos usa.
¿Porque?: Conocer los modelos con los que se ayudan las empresas para realizar el análisis de su información.
RE: Selección de alguno de los modelos presentados.

6. ¿Los modelos que usa le permiten hacer análisis exhaustivos de su información histórica de manera eficiente?
¿Porqué?: Conocer el nivel de detalle de análisis que les proporcionan los modelos que utilizan.
RE: No.

7 ¿Usa herramientas automáticas orientadas a la generación de información que le ayuden en la toma de decisiones?
¿Porqué?: Permite identificar si se apoya en herramientas automáticas de análisis de información para su toma de decisiones.
RE: No.

8. Si contestó afirmativamente la pregunta anterior, marque con una "X" qué herramientas usa.
¿Porqué?: Conocer en que herramientas automáticas se apoyan los directivos para la toma de decisiones.
RE: Selección de algunas de las herramientas automáticas de apoyo para la toma de decisiones presentadas.

9. ¿Considera que la información generada por sus sistemas actuales le permite conocer a detalle su negocio?
¿Porqué?: Conocer el alcance de las herramientas que usan actualmente los directivos y medir la capacidad de análisis de la información histórica generada por éstos.
RE: No.

10. ¿Considera que el utilizar herramientas automáticas que le permitan generar información adicional a la que obtiene de sus sistemas, le puede ayudar a incrementar la productividad de su empresa?
¿Porqué?: Demostrar la importancia que tiene el uso de las herramientas de análisis de información histórica para explotar la información y generar conocimiento.
RE: Si.

11. ¿Conoce el término "minería de datos"?
¿Porqué?: Es nuestro tema de investigación y se desea saber que tan familiarizados están los expertos en la materia con este término.
RE: No.

1.4.1 Lista de personas entrevistadas.

Nombre	Puesto	Dirección
Ing. Alejandro Gabino Gallardo e-mail: agabino@ford.com	Jefe de Departamento	Ford Motor Company
Ing. Gabriela Betzabé Lizárraga Ramírez	Ingeniero en Sistemas	Facultad de Ingeniería - UNAM
Ing. Martín de Jesús Jiménez e-mail: martin@drbaz.fmedic.unam.mx	Jefe de Unidad de Sistemas	Facultad de Medicina - UNAM
Lic. Humberto Hernández López	Ingeniero en Sistemas	Banamex
Lic. Jorge Eduardo González Cuellar e-mail: Jorge@drbaz.fmedic.unam.mx	Jefe de Departamento	Facultad de Medicina - UNAM
Lic. Lourdes Aida Granados Granados	Ingeniero en Sistemas	Banamex

1.4.2 Conglomerado y análisis de las respuestas a los cuestionarios aplicados.

A continuación se resumen y analizan las respuestas proporcionadas por las personas encuestadas.

1. ¿En la organización en la que labora, utilizan algún modelo que les ayude al diseño de la información?

Si ( ) No ( )

Análisis: A través de las respuestas obtenidas se aprecia que todos los usuarios encuestados si hacen uso de modelos para el diseño de información.

Conclusión: Según las respuestas otorgadas todos los encuestados hacen uso de algún modelo para el análisis de su información.

2. Si contestó afirmativamente la pregunta anterior, señale con "X" que modelos usa.

( ) Modelo relacional ( ) Modelo jerárquico
( ) Modelo de red ( ) Modelo distribuido
Otros:

Análisis: Los seis usuarios entrevistados trabajan con el modelo de bases de datos relacional.

Conclusión: De acuerdo a las respuestas de los encuestados todos trabajan con información normalizada, organizada y relacionada; lo que facilita, enormemente, el trabajo a los programas mineros en el proceso de la minería de datos.

3. ¿En la organización en la que labora, diseñan las bases de datos para su explotación inmediata o futura?

Inmediata ( ) Futura ( ) Ambas ( )

Análisis: Sólo una persona diseña sus bases de datos para su explotación inmediata. Dos diseñan sus bases para su explotación futura y tres diseñan sus bases para su explotación inmediata y futura (ambas).

Conclusión: La persona que diseña las bases de datos de forma inmediata sólo atiende o satisface necesidades rutinarias, siendo mínimo el aprovechamiento y explotación de las bases de datos.
Por otro lado, las dos personas que diseñan las bases de datos para su explotación futura tienen una visión de las posibles necesidades que pudieran darse, pero no están contemplando las necesidades actuales y eso hace que no se tomen en cuenta aspectos trascendentales de las necesidades en el diseño de las bases de datos.
Por último las tres personas que diseñan las bases de datos para su explotación inmediata y futura tienen un panorama general de lo que son sus necesidades actuales y futuras; por lo que su diseño e implementación de las bases de datos ayudará a una mejor explotación y aprovechamiento.
De acuerdo a las respuestas obtenidas nos permiten ver que actualmente los usuarios tienen una visión mas completa en el diseño de las bases de datos.

4.¿Usa modelos que le ayuden al análisis de su información histórica?

Si ( ) No ( )

Análisis: De las cinco personas entrevistadas ninguna usa modelos para el análisis de su información histórica.

Conclusión: Como se conjetura en nuestra hipótesis, actualmente ninguna de las personas se apoyan en modelos que le ayuden al análisis de su información histórica.

5. Si contestó afirmativamente la pregunta anterior, señale con "X" qué modelos usa.

( ) Clasificación
( ) Asociación
( ) Árboles de decisiones
( ) Redes neuronales
( ) Algoritmos paralelos
( ) Híbrido (Basado en arquitectura de redes neuronales)
( ) Reglas de inducción
( ) Estadísticas
( ) Análisis fractal
( ) Clustering
( ) Sequencias
( ) Algoritmos genéticos
( ) Memoria basada en razonamiento
( ) Reglas de clasificación
( ) Patrones generales y búsqueda de excepciones
( ) Sistemas de visualización
( ) Técnicas difusas
( ) Sistemas de información geográfica

Análisis: Ninguno de las cinco personas entrevistadas selecciono uno de los modelos para el análisis de la información histórica.

Conclusión: Como se esperaba, ninguna de las personas hacen uso de los modelos mencionados en la lista para el análisis de su información.

6. ¿Los modelos que usa le permiten hacer análisis exhaustivos de su información histórica de manera eficiente?

Sí ( ) No ( )

Análisis: De acuerdo con las respuestas de las cinco personas entrevistadas, ninguna hace uso de modelos que les permitan hacer análisis exhaustivos de su información.

Conclusión: Nuevamente se refleja el desaprovechamiento de la información histórica; ya que ninguna de las personas hace ningún tipo de análisis de su información acumulada.

7. ¿Usa herramientas automáticas orientadas a la generación de información que le ayuden en la toma de decisiones?

Sí ( ) No ( )

Análisis: De las personas entrevistadas, todas contestaron que no usan herramientas automáticas orientadas a la generación de información que les ayude a la toma de decisiones.

Conclusión: Se deduce que estas personas se limitan a que sus decisiones sean soportadas sólo por los sistemas operacionales y específicos de su negocio.

8. Si contestó afirmativamente la pregunta anterior, marque con una "X" qué herramientas usa.

( ) DataCruncher
( ) DataMind
( ) AgentBase
( ) Business Miner
( ) DataBase Mining Marksman
( ) DataEngine
( ) MineSet
( ) Relationship Manager
( ) DecisionMaster
( ) Decision Series
( ) Ultragem Data Mining
( ) SuperQuery
( ) ModelMAX
( ) Data Surveyor
( ) Alice, AC2
( ) Management Discovery Tool (MDT)
( ) DataDetective
( ) Syllogic Datamining Tool/MP
( ) Decisionhouse
( ) Otras:

Análisis: Para efectos de este análisis, la pregunta queda sin valor ya que de acuerdo a la respuesta anterior ninguna de las personas entrevistadas hace uso de herramientas para el soporte a la toma de decisiones.

Conclusión: Se confirma que ésta lista de herramientas no son aplicadas por las personas entrevistadas.

9. ¿Considera que la información generada por sus sistemas actuales le permite conocer a detalle su negocio?

Sí ( ) No ( )

Análisis: Dos de las personas encuestadas opinan que sus sistemas son suficientes para conocer a detalle su negocio. Tres de ellos opinan lo contrario.

Conclusión: Aunque solo fueron cuatro las personas que apoyan nuestra respuesta esperada se sigue demostrando que la información obtenida por sus sistemas actuales les es insuficiente para conocer a detalle su negocio.

10. ¿Considera que el utilizar herramientas que le permitan generar información adicional a la que obtiene de sus sistemas, le puede ayudar a incrementar la productividad de su empresa?

Sí ( ) No ( )

Análisis: De acuerdo con las opiniones de los entrevistados todos señalan la importancia del uso de herramientas que ayuden a incrementar la productividad de la empresa.

Conclusión: Se ve claramente la importancia y necesidad de conocer y usar herramientas de análisis de información histórica para explotar la información y generar conocimiento.

11. ¿Conoce el término "minería de datos"?

Sí ( ) No ( )

Análisis: Sólo una persona no conoce el término de minería de datos, las otras cuatro sí están familiarizados con el término.

Conclusión: Es importante saber que las personas están familiarizadas con el tema y que no les es novedad.

1.4.3 Conclusión general

El 81.81% de las personas entrevistadas están conscientes de que sus sistemas actuales no les permiten conocer a detalle su negocio, de que no cuentan con herramientas automáticas que les permitan hacer análisis detallados de su información y de que existe un gran desaprovechamiento de su información histórica. Todos ellos, al igual que nosotras, opinan que el contar con herramientas que les permitan generar

información adicional, para el soporte a la toma de decisiones, les ayudaría a incrementar la productividad de su empresa.

1.5 Opiniones profesionales

Se espera recopilar la opinión y punto de vista de personas profesionales con cierto grado de estudios y experiencia en el área de informática con objeto de sustentar el problema identificado y demarcado en nuestro trabajo.

Para tal efecto se aplicó el mismo cuestionario diseñado en el punto 1.4.

1.5.1 Lista de personas entrevistadas:

Nombre	Puesto	Ubicación
Ing. Julian A. Juárez Hernández e-mail: **¡Error! Marcador no definido.**	Coordinador de Help Desk	GE Capital Information Technology Solutions
Ing Alexei Samuel Dezotti Ruiz	Jefe del Departamento de Innovación de la Subdirección de Sistemas	Dirección de Cómputo Académido para la Administración - UNAM
Dr. Adolfo Guzmán Arenas	Director	Centro de Investigación en Computación - IPN
Ing. Sergio Rivera Romero	Líder de proyecto	Dirección General de Administración Escolar - UNAM
Lic. Ma. Antonieta Reza-Garduño O.	Subdirectora de Productividad y Nueva Tecnologia	Banpaís
Ing. Armando Vega A.	Jefe de Unidad	Dirección General de Administración Escolar - UNAM

1.5.2 Conglomerado y análisis de las respuestas a los cuestionarios aplicados

A continuación se resumen y analizan las respuestas proporcionadas por las personas encuestadas.

1. ¿En la organización en la que labora, utilizan algún modelo que les ayude al diseño de la información?

Si ( ) No ( )

Análisis: A través de las respuestas obtenidas se aprecia que todos los usuarios encuestados sí hacen uso de modelos para el diseño de información.

Conclusión: Según las respuestas otorgadas todos los encuestados hacen uso de algún modelo para el análisis de su información.

2. Si contestó afirmativamente la pregunta anterior, señale con "X" que modelos usa.

( ) Modelo relacional ( ) Modelo jerárquico
( ) Modelo de red ( ) Modelo distribuido
Otros:

Análisis: Los seis usuarios entrevistados trabajan con el modelo de bases de datos relacional.

Conclusión: De acuerdo a las respuestas de los encuestados todos trabajan con información normalizada, organizada y relacionada; lo que facilita, enormemente, el trabajo a los programas mineros en el proceso de la minería de datos.

3. ¿En la organización en la que labora, diseñan las bases de datos para su explotación inmediata o futura?

Inmediata ( ) Futura ( ) Ambas ( )

Análisis: Sólo dos personas diseñan sus bases de datos para su explotación inmediata y cuatro diseñan sus bases para su explotación inmediata y futura (ambas).

Conclusión: Las personas que diseñan las bases de datos de forma inmediata sólo atienden o satisfacen necesidades rutinarias, siendo mínimo el aprovechamiento y explotación de las bases de datos.

Las personas que diseñan las bases de datos para su explotación inmediata y futura tienen una visión general de lo que son sus necesidades actuales y futuras; por lo que su diseño e implementación de las bases de datos ayudará a una mejor explotación y aprovechamiento de la información.

De acuerdo a las respuestas obtenidas nos permiten ver que actualmente los usuarios tienen una visión mas completa en el diseño de las bases de datos.

4. ¿Usa modelos que le ayuden al análisis de su información histórica?

Sí ( ) No ( )

Análisis: Cinco personas entrevistadas hacen uso de modelos para el análisis de su información histórica, incluso hacen una combinación de éstos. Sólo una no hace uso de herramientas para el análisis de su información.

Conclusión: Los expertos en la materia se apoyan en herramientas para el análisis de su información histórica.

5. Si contestó afirmativamente la pregunta anterior, señale con "X" que modelos usa.

( ) Clasificación
( ) Asociación
( ) Árboles de decisiones
( ) Redes neuronales
( ) Algoritmos paralelos
( ) Híbrido (Basado en arquitectura de redes neuronales)
( ) Reglas de inducción
( ) Estadísticas
( ) Análisis fractal
( ) Clustering
( ) Sequencias
( ) Algoritmos genéticos
( ) Memoria basada en razonamiento
( ) Reglas de clasificación
( ) Patrones generales y búsqueda de excepciones
( ) Sistemas de visualización
( ) Técnicas difusas
( ) Sistemas de información geográfica
Otras:

Análisis: Cuatro de las de los entrevistados usan el modelo de clasificación, uno usa el modelo de asociación, tres usan arboles binarios, uno usa redes neuronales, cuatro usan modelos estadisticos, uno usa cluster, uno usa algoritmos genéticos, uno usa reglas de clasificación y dos usan patrones generales y búsqueda de excepciones.

Conclusión: Al menos uno de los modelos presentados es conocido y aplicado por la mayoría de las personas entrevistadas.

6. ¿Los modelos que usa le permiten hacer análisis exhaustivos de su información histórica de manera eficiente?

Sí ( ) No ( )

Análisis: De acuerdo con las respuestas de las seis personas entrevistadas, cuatro opinan que si y dos que no.

Conclusión: Los expertos, aún con el uso de modelos, no están completamente satisfechos con los análisis obtenidos en su información histórica. Por lo que es necesario hacer análisis más profundos con la minería de datos.

7. ¿Usa herramientas automáticas orientadas a la generación de información que le ayuden en la toma de decisiones?

Sí ( ) No ( )

Análisis: Tres de las seis personas entrevistadas señalan el uso de herramientas para la generación de información que apoye la toma de decisiones. Tres de ellas señalan que no hacen uso de herramientas.

Conclusión: No todas las personas que se apoyan en modelos de análisis cuentan con herramientas para la explotación de la información.

8. Si contestó afirmativamente la pregunta anterior, marque con una "X" qué herramientas usa.

( ) DataCruncher
( ) DataMind
( ) AgentBase
( ) Business Miner
( ) DataBase Mining Marksman
( ) DataEngine
( ) MineSet
( ) Relationship Manager
( ) DecisionMaster
( ) Decision Series
( ) Ultragem Data Mining
( ) SuperQuery
( ) ModelMAX
( ) Data Surveyor
( ) Alice, AC2
( ) Management Discovery Tool (MDT)
( ) DataDetective
( ) Syllogic Datamining Tool/MP
( ) Decisionhouse
( ) Otras:

Análisis: Sólo dos de las personas hacen uso de herramientas para la explotación de la información.

Conclusión: Una de las herramientas es DecisionMaster y Anayne Gux(OLAP). La mayoría de los expertos entrevistados no hacen uso de ninguna herramienta de análisis.

9. ¿Considera que la información generada por sus sistemas actuales le permite conocer a detalle su negocio?

Sí ( ) No ( )

Análisis: Tres de las personas encuestadas consideran que sus sistemas son los adecuados para conocer a detalle su negocio y tres opinan lo contrario.

Conclusión: La información generada por los sistemas operacionales es suficiente para conocer a detalle la empresa de tres de los encuestados, sin embargo, en la pregunta diez reconocen la importancia de obtener información adicional que les ayuda a incrementar la productividad de la empresa.

10. ¿Considera que el utilizar herramientas que le permitan generar información adicional a la que obtiene de sus sistemas, le puede ayudar a incrementar la productividad de su empresa?

Sí ( ) No ( )

Análisis: De acuerdo con las opiniones de los entrevistados todos señalan la importancia del uso de herramientas que ayuden a incrementar la productividad de la empresa.

Conclusión: Los expertos entrevistados admiten que el uso de herramientas que les permitan generar información adicional a la que obtienen de sus sistemas les ayudaría a incrementar la productividad de la empresa.

11. ¿Conoce el término "minería de datos"?

Sí ( ) No ( )

Análisis: Todas las personas están familiarizadas con el término minería de datos.

Conclusión: Es bueno saber que la opinión de los expertos coincide con nuestra problemática identificada en la hipótesis.

1.5.3 Conclusión general

El 65.15% de las respuestas de los expertos entrevistados aseveran que si utilizaran una herramienta de análisis de información podrían conocer más a detalle su negocio, lo que les permitiría tomar decisiones más acertadas para incrementar la producctividad de la empresa.

1.6 Hipótesis preliminar

1.6.1 Presentación del problema en su relación causa-efecto

VARIABLES INDEPENDIENTES	VARIABLES DEPENDIENTES y/o INDEPENDIENTES	VARIABLES DEPENDIENTES
• Gran cantidad de información.	• Acumulación sin sentido de la información. • Uso inadecuado de la información. • Mala organización • Pérdida de información • Inconsistencia en la información. • Es difícil obtener información confiable. • Desconocimiento de la información histórica.	
	• La información que es obtenida no es confiable	• las decisiones que se toman no son certeras.
• Muchas situaciones interesantes se soslayan (no se detectan).	• Toma mucho tiempo pensar o diseñar una consulta, elaborarla, expresarla e interpretarla.	

• Es complicado identificar [1]hallazgos.	• Análisis complicados y tardados. • Fuertes inversiones de personal, tiempo y dinero.	
• Es complicado definir criterios de selección de información que ayuden en el descubrimiento de situaciones interesantes.	• Análisis complicados y tardados. • Fuertes inversiones de personal, tiempo y dinero.	
• El no uso de métodos y criterios de selección o de búsqueda de datos importantes para la empresa.	• Incapacidad para descubrir patrones de comportamiento en los datos que ayuden a detectar fallas o errores en los mismos, de manera rápida.	
• El no uso de herramientas de análisis automático de información.	• Alto costo de oportunidad. El costo por no haber invertido en algo que pudo ofrecer oportunidades. • Decremento de la competitividad en las empresas. • Desconocimiento de: tendencias, gustos, preferencias y comportamiento del mercado en el futuro	

Correlación de variables dependientes e independientes

1. La gran cantidad de información hace que la misma sea acumulada sin sentido.
2. La gran cantidad de información provoca su uso inadecuado.
3. La gran cantidad de información provoca su mala organización.
4. La gran cantidad de información provoca pérdida de la misma.
5. La gran cantidad de información provoca inconsistencia.
6. La gran cantidad de información provoca desconocimiento de la información histórica.

[1] Descubrimiento o invención de algún hecho u obra.

7. La gran cantidad de información hace difícil la obtención de información confiable.

8. La obtención de información no confiable hace que las decisiones que se toman no sean certeras.

9. Muchas situaciones interesantes se soslayan (no se detectan) debido a que toma mucho tiempo pensar o diseñar una consulta , elaborarla, expresarla e interpretarla.

10. Es complicado identificar hallazgos ya que es necesario hacer análisis complicados y tardados.

11. Es complicado identificar hallazgos por lo que se requiere de fuertes inversiones de personal, tiempo y dinero.

12. El no uso de métodos y criterios de selección o de búsqueda de datos importantes para la empresa provoca incapacidad para descubrir patrones de comportamiento en los datos que ayuden a detectar fallas o errores en los mismos, de manera rápida.

13. El no uso de herramientas de análisis automático de información trae consigo un alto costo de oportunidad.

14. El no uso de herramientas de análisis automático de información frena la competitividad en las empresas.

15. El no uso de herramientas de análisis automático de información impide el descubrimiento de tendencias, gustos, preferencias y comportamiento del mercado en el futuro.

1.6.2 Ejemplos de correlación de hipótesis

No altera

1. La gran cantidad de información hace difícil la obtención de información confiable.
   La poca cantidad de información hace fácil la obtención de información no necesariamente confiable.

2. La obtención de información no confiable hace que las decisiones que se toman no sean certeras.
   La obtención de información confiable hace que las decisiones que se toman sean certeras.

3. Muchas situaciones interesantes se soslayan (no se detectan) debido a que toma mucho tiempo pensar o diseñar una consulta , elaborarla, expresarla e interpretarla.
   Muchas situaciones interesantes se detectan debido a que toma poco tiempo

pensar o diseñar una consulta , elaborarla, expresarla e interpretarla.

4. Es complicado identificar hallazgos ya que es necesario hacer análisis complicados y tardados.
   Es fácil identificar hallazgos por lo que no es necesario hacer análisis complicados y tardados.

5. Es complicado identificar hallazgos por lo que se requiere de fuertes inversiones de personal, tiempo y dinero.
   Es sencillo identificar hallazgos por lo que no se requiere de fuertes inversiones de personal, tiempo y dinero.

6. El no uso de métodos y criterios de selección o de búsqueda de datos, importantes para la empresa, provoca incapacidad para descubrir patrones de comportamiento en los datos que ayuden a detectar fallas o errores en los mismos, de manera rápida.
   El no uso de métodos y criterios de selección o de búsqueda de datos, importantes para la empresa, permite descubrir patrones de comportamiento en los datos que ayuden a detectar fallas o errores en los mismos, de manera rápida.

7. El no uso de herramientas de análisis automático de información trae consigo un alto costo de oportunidad.
   El uso de herramientas de análisis automático de información trae consigo un bajo costo de oportunidad.

8. El no uso de herramientas de análisis automático de información impide el descubrimiento de tendencias, gustos, preferencias y comportamiento del mercado en el futuro.
   El uso de herramientas de análisis automático de información permite el descubrimiento de tendencias, gustos, preferencias y comportamiento del mercado en el futuro.

Si altera

1. La gran cantidad de información hace que la misma sea acumulada sin sentido.
   La poca cantidad de información hace que la misma sea acumulada con sentido.

2. La gran cantidad de información provoca su uso inadecuado.
   La poca cantidad de información provoca su uso adecuado.

3. La gran cantidad de información provoca su mala organización.
   La poca cantidad de información provoca su buena organización.

4. La gran cantidad de información provoca pérdida de la misma.
   La poca cantidad de información no provoca pérdida de la misma.

5. La gran cantidad de información provoca inconsistencia.
   La poca cantidad de información no provoca inconsistencia.

6. La gran cantidad de información provoca desconocimiento de la información histórica.
   La poca cantidad de información provoca desconocimiento de la información histórica.

7. El no uso de herramientas de análisis automático de información frena la competitividad en las empresas.
   El uso de herramientas de análisis automático de información acelera la competitividad en las empresas.

1.6.3 Hipótesis preliminar

En la mayoría de las empresas que manejan grandes volúmenes de información, la falta de conocimiento y la poca o nula utilización de herramientas automáticas que permitan el análisis profundo de información histórica, les impide definir criterios de identificación, selección y búsqueda de datos importantes que les ayuden a conocer más a detalle su negocio y a tomar mejores decisiones.

1.7 Objetivos

1.7.1 Personales

María Isabel Angeles Larrieta

- Concluir mis estudios a través de una tesis.
- Obtener el grado de Licenciatura y poder continuar con una maestría.
- Proyectarme como una futura líder de proyectos.

Angélica María Santillán Gómez

- Culminar con los esfuerzos realizados durante 18 años de estudio continuo y obtener un título que me acredite como profesional.

María Isabel Angeles Larrieta y Angélica María Santillán Gómez

- Cumplir con los requisitos marcados por el Reglamento de Exámenes Profesionales en sus artículos 19 y 20 en donde se señala que se expedirá el título a quienes hayan aprobado el examen profesional oral y escrito siendo éste último, la elaboración de un trabajo de investigación (tesis).
- Obtener el título de Lic. en Informática otorgado por la máxima casa de estudios, la Universidad Nacional Autónoma de México.

1.7.2 Particulares

María Isabel Angeles Larrieta

- Volverme experta en la minería de datos, considero que es una gran herramienta de apoyo para cualquier empresa.
- Participar como agente de cambio en la cultura de información estratégica de la compañía donde labore.
- Escribir acerca del tema en revistas y periódicos especializados en informática para promover los beneficios resultantes de la buena aplicación de esta metodología de explotación de información.
- Participar, como parte del grupo de trabajo, en proyectos relacionados con la minería de datos.

Angélica María Santillán Gómez

- Reunir una gran cantidad de conocimientos sobre el tema minería de datos y así volverme especialista y proporcionar mis servicios a empresas que lo requieran, obteniendo de esta manera beneficios mutuos.

María Isabel Angeles Larrieta y Angélica María Santillán Gómez

- Volvernos expertas en el tema para ofrecer consultoría.

1.7.3 Generales

- Dar a conocer la minería de datos como un proceso para la explotación de la información histórica generada por una empresa.
- Hacer un trabajo de investigación que ayude a todas las empresas a explotar mejor su información.

Marco teórico

2. MARCO TEORICO

2.1 Acopio bibliográfico

Con el objetivo de tener un mayor conocimiento y manejo del tema, nos abocamos en la tarea de buscar y consultar las fuentes de información que nos proporcionaron diferentes perspectivas en el proceso de minería de datos.

A continuación se mencionan los diversos tipos de lecturas consideradas:

Lectura total.- Es aquella que se realiza de todo el libro.
Lectura ligera.- Es la consulta de uno o varios capítulos.
Lectura rápida.- Es la referente a un tema en específico.
Lectura superficial.- Es aquella de uno o varios conceptos.

2.1.1. Libros

Consulta total	
BERRY J.A., Michael y Gordon **Linoff**, Data Mining Techniques for Marketing, Sales and Customer Support, Ed. Willey Computer Publishing, 1997 United States.	Se consultó todo el libro.
Consulta Ligera	
ABITEBOUL, SERGE, Richard **Hull** and Victor **Vianu**, Foundations of databases, Ed. Addison-Wesley, 1a. ed. 1995.	El capítulo que se consultó fue "Evolution of the database", habla de la evolución de las bases de datos, su estructura y organización.
ADAD, Rubén, Alfredo **Careaga**, Migel Ángel **Medina**, Fundamentos de bases de datos relacionales, Grupo Noriega Editores,1° ed. 1993.	Los capítulos que se consultaron son: Modelos de bases de datos, el modelo de base de datos relacional y SQL.
CARDENAS, Alfonso F., Sistemas de administración de bases de datos. Limusa, México 1985.	El capítulo que se consultó fue "Manejo y organización de archivos", habla de la evolución, estructura, organización y formas de acceso de los archivos.
HARJINDER S., Gill y Rao **Prakash C**, Data warehousing: La integración de información para la mejor toma de decisiones, ed. Prentice Hall Hispanoamericana S. A., 1ª edición 1996. Pp. 239-261	El capítulo que se consultó fue "Minería de Datos", habla de la forma en que la minería de datos ayuda a los usuarios empresariales a procesar grandes cantidades de información, y de las características de la minería de datos.
MARK L. Gillenson, Introducción a las bases de datos, IBM System Research Institute, McGraw-Hill, 1988 ISBN 968-422-303-X	El capítulo que se consultó fue "Orígenes de los registros", da una introducción de la evolución de los datos y archivos; así como también, habla de los conceptos de almacenamiento.
SALEMI, Joe, PC Magazine Guide to Client/Server Databases, Emeryville, California: Ziff Davis, 1995, Pp. 312	El capítulo que se consultó fue "Architecture Client/Server", habla de la estructura y tecnologías del sistema cliente/servidor, muestra las ventajas y beneficios de dicho sistema.

2.1.1.1 Diccionarios

De la Lengua Española	
Diccionario de la Lengua Española **Real Academia Española** España - Calpe, S.A. 19. Edición.	Se consultaron los términos: dato, minería de datos.
De Sinónimos y Antónimos	
Larouse 1 **Conjugación, Sinónimos y Antónimos** Ediciones Larouse, Editora de Periódicos, S.C.L., México 1992.	Se consultaron los sinónimos de datos y minería de datos.
De Informática	
Diccionario de Informática Trad. Blanca de Mendizabal Allende, Editorial Díaz de Santos, México 1990. ISBN 84-7968-068-1	Se consultaron términos informáticos para agregarlos al glosario.
Diccionario de Minicomputadores y microcomputadores **PHILIP E,** Burton, De. URMO, S.A. Edición en español. ISBN 84-314-0451-5	Se consultaron términos informáticos para agregarlos al glosario.
Glosario de términos y siglas **Diccionario Inglés - Español** **SÁNCHEZ VAQUERO,** Antonio Luis **Joyanes Aguilar,** Mc GrawHill México 1985. ISBN 968-451-785-0	Se consultaron términos informáticos para agregarlos al glosario.

2.1.2 Tesis

El único trabajo de investigación desarrollado hasta el momento y relacionado con nuestro tema es el de Data warehousing.

Consulta ligera	
RUIZ TORRES, Mary Karina y Ney **Galicia Arrocena,** Data warehousing como factor competitivo en la toma de decisiones. Tesis de Licenciatura (en informática), México: UNAM-FCA, 1997, 152 p.	El capítulo que se consultó fue "Minado de datos", habla de las aplicaciones, enfoques, algoritmos y modelos de la minería de datos.

2.1.3 Revistas

Consulta ligera	
GUZMÁN ARENAS, Adolfo, Uso y diseño de mineros de datos, Soluciones Avanzadas, Año 4, Número 34, Junio de 1996, pp 67-72.	Se consulto el artículo "**Uso y diseño de mineros de datos**". El Dr. Adolfo Guzmán Arenas hace una descripción general de los mineros de datos, presenta las ventajas y desventajas de los mineros, así como su estructura y diseño.
CHERYL D., Krivda, Unearthing Underground Data, Data mining, Vol. 11, Número 5, Mayo de 1996.	Se consulto el articulo "**Unearthing Underground Data**". Cheryl D. Krivda hace un resumen de los mineros y sus características.
Consulta total	
IBM, Datamining y Datawarehousing, Informe de IBM Software Update Fecha: 1997	Se consultó todo el informe.
IBM, Gestión competitiva de la información, Informe de IBM Software Update Fecha: 1997	Se consultó todo el informe.

2.1.4 Periódicos

Consulta ligera	
COMPUTERWORLD, Data Mining, a pesar de los peligros, Año 18, Número 553, Febrero 16-20 1998.	Habla de lo peligroso que tiene aplicar la minería de datos, y de su proyección según el Gartner Group para el futuro.

2.1.5 Seminarios

Fecha	Título	Expositor
10 May 96	On modeling the degree of classification error reduction obtained by combining classifiers, **Data Mining Solutions**	Kamal Ali
17 May 96	**SLIQ: A Fast Scalable Classifier for Data Mining**	Manish Mehta
24 May 96	**Mining Sequential Patterns**	Ramakrishnan Srikant
31 May 96	**Fast Serial and Parallel Classification of Very Large Data Bases**	John C. Shafer
07 Jun 96	**Flexibly Exploiting Prior Knowledge in Empirical Learning**	Julio Ortega, Data Mining Solutions.
14 Jun 96	**Building Classifiers using Bayesian Networks**	Moises Goldszmidt, SRI International
21 Jun 96	**Learning Networks Applied to Pattern Recognition**	Ying Zhou, Data Mining Solutions.
28 Jun 96	**Concept Induction and Temporal Data**	Stefanos Manganaris, Data Mining Solutions
12 July 96	**The WoRLD: Knowledge Discovery from Multiple Distributed Databases**	Venkat Kolluri, Data Mining Solutions
26 July 96	**MDL estimation for small sample sizes and its application to linear regression**	Byron E. Dom
4 October 96	**Mining quantitative association rules**	Ramakrishnan Srikant
01 November 96	**Discovering Trends in Text Databases**	Brian Lent
18 Jun 97	**Parallel Association Rules**, Mohammed Zaki	IBM Almaden (also University of Rochester).
26 Jun 97 (11 am.)	**Storage Management and Data Mining problems in High Energy Physics application**	Arie Shoshani, Lawrence Berkeley National Laboratory, Berkeley, California.
2 July 97	**Computing Most Specific Sentences that are Interesting in a Database**	Dimitrios Gunopulos, IBM Almaden
16 July 97 (2 pm.)	**High-Dimensional Similarity Join**	Ramakrishnan Srikant, IBM Almaden
29 July 97 (11 am.)	**Mining for Associations over Interval Data**	Renee J. Miller, Ohio State University.
30 July 97 (11 am.)	**Query Flocks: A Framework for Large-Scale Data Mining**	Jeffrey D. Ullman, Stanford University
1 August 97 (11 am.)	**Performance Optimization in the Knowledge Discovery in Database -Process**	Peter Lockeman, University of Karlsruhe, Germany,
17 September 97, (2 pm.)	**Learning Relational Rules from Relational Data**	Kamal Ali, IBM Global Business Intelligence Solutions
17 October 97 (11 am.)	**A Microeconomic View of Data Mining**	Christos Papadimitriou, UC Berkeley.
17 October 97 (3 p.m.-5 p.m.)	**Tutorial on Neural Networks for Pattern Recognition, Session I**	Jianchang Mao, IBM Almaden.
31 October 97	**Tutorial on Neural Networks for Pattern**	Jianchang Mao, IBM

(2:30 p.m.-4:30 p m)	**Recognition, Session II**	Almaden.
5 November 97 (11 a.m.)	**Mining the Web**	Sergey Brin, Stanford
18 November 97 (2 pm)	**Discovery of Patterns in Very Large Dimension Data Sets using Hypergraph Models**	Vipin Kumar, University of Minnesota.

2.1.6 Conferencias

Fecha	Titulo	Lugar
November 1 - 5, 1993	**Call for papers: Knowledge discovery in scientific databases**, A Special Session of the Second International Conference on Information and Knowledge Management..	WashingtonD.C., USA local copy (ASCII)
October 12-15, 1993	**CFP for International Workshop on Rough Sets and Knowledge Discovery (RSKD-93)** To be held at Banff, Alberta, Canada	
July 31 August1, 1994	**CFP for KDD-94: AAAI Workshop on Knowledge Discovery in Databases**	Seattle, Washington
Wednesday 1st of Februrary 1995	**IEE / BCS Colloquium on KDD, Location : IEE**	London.Organised by the Institute of Electrical Engineers professional group C4 - Atrificial Intelligence and the British Computer Society Specialist Group on Expert Systems (SGES)
March 2, 1995	**CSC'95 Workshop on Rough Sets and Database Mining**	Location: Nashville, Tennessee, USA
August 20-21, 1995	**First International Conference on Knowledge Discovery and Data Mining**	Location: Montreal, Canada Note: Co-located with IJCAI-95
September 11 –15, 1995	**21st International Conference on Very Large Data Bases 1995 (VLDB'95)**	Location: Zurich, Switzerland, Note: Includes 2 sessions on data mining in the science track.
September 17-18, 1995	**INTERNATIONAL WORKSHOP ON TEMPORAL DATABASES**	Location: Zurich, SWITZERLAND
December 8, 1995	**First International Workshop on the Integration of Knowledge Discovery with Deductive and Object-Oriented Databases (KDOOD)**	
April 28-29, 1996	**Knowledge Level Modelling and Machine Learning Workshop**	Location: Iraklion, Greece
Jun2, 1996	**Workshop on Research Issues on Data Mining and Knowledge Discovery**	Location: Montreal, Canada
July 3-6, 1996	**13th International Conference on Machine Learning**	Location: Bari, Italy
July 1996	**Data Mining 96**	Location: Sydney, Australia, Contributed by: Steve Hitchman
August 3-5, 1996	**The Second International Conference on Knowledge Discovery and Data Mining (KDD-96)**	Location: Portland, Oregon, USA

February 24-27, 1997	**First Pacific-Asia Conference on Knowledge Discovery and Data Mining**	Location: Singapore
April 26th, 1997	**International Conference on Ordinal and Symbolic Data Analysis**	Location: Prague, Czech Republic , Contributed by: charles@amsta.leeds.ac.uk
August 4-6, 1997	**Second International Symposium on Intelligent Data Analysis**	Location: Birkbeck College, London, England

2.1.7 Internet

URL	Artículo
http://www-pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_2.html	**Data mining**, What is data mining? For *William J Frawley, Gregory Piatetsky-Shapiro and Christopher J Matheus and Marcel Holshemier & Arno Siebes* (1994).
http://pwp.starnetinc.com/larryg/books.html	**Books on Data Mining and Decision Support**
http://www.almaden.ibm.com/cs/quest/seminars-hist.html	**Data Mining Semminars**
http://www.almaden.ibm.com/cs/quest/publications.html	**Quest Publications**
http://www.pilotsw.com/whtpaper/datamining/dmarch.htm	**An Architecture for Data Mining**
http://www.gslis.utexas.edu/~palmquis/courses/project/dm_basic.htm	**Popular Data Mining Methods**
http://www.datamining.com/satamine/powerful.htm	**What Makes the Data Mining Suite So Powerful?**
http://www.ddi.nl/pub/whitepaper.htm_Toc383226494	**On Line Data Mining, Data Distilleries vision on Data Mining technology**
http://marvin.cs.uah.edu/~jrushing/mining.html	**Bibliografía**
http://shop.barnesandnoble.com/BookSearch/results.asp	**Bibliografía, Última actualización. 24 de Septiembre de 1997**
http://ciir.cs.umass.edu/info/query_syntax_demoOld.html	**Query Formulation**
http://users.demag.rwth-aachen.de/donald/Diversen/W3Encyc/_quer003.htm	**Query**
http://rcr.csun.edu/ChrisJ/computers/sql/sqlch3.htm	**Relational Databases**
http://www.cs.bham.ac.uk/~anp/conferences.html	**Data Mining Conferences**
http://www-pcc.qub.ac.uk./tec/courses/datamining/stu_notes/dm_book_2.html	**Data Mining para William J. Frawley, Gregory Piateski-Shapiro y Christopher J. Matheus**

2.1.8 Proveedores

Compañía	URL	Software
AbTech Corporation	http://www.abtech.com/	*ModelQuest*
AcknoSoft	http://www.acknosoft.com	*KATE*
Advanced Software Applications Corp	http://www.asacorp.com	*ModelMAX*
Attar Software	http://www.attar.com	*XpertRule Profiler*
Automatic Forecasting Systems, Inc.	http://darkstar.icdc.com/~autobox	*Autobox*
AZMY Thinkware Inc.	http://www.azmy.com	*SuperQuery*
Business Objects	http://www.businessobjects.com	*Business Miner*
Cap Gemini bv	http://www.capgemini.nl	*Omega*
Cognos	http://www.cognos.com	*Scenerio*
Data Distilleries	http://www.ddi.nl	*Data Surveyor*
DataMind Corp	http://www.datamind.com	*DataMind*
Datasage	http://www.datasage.com	*datasage*
DAZ Systems Inc	http://www.dazsi.com	*AgentBase*
HNC Software Inc	http://www.hnc.com	*DataBase Mining Marksman*
Hyperparallel	http://www.hyperparallel.com	*Discovery*
Information Discovery	http://www.datamine.inter.net	*IDIS*
Integral Decisions Systems SE GmbH (INDECS)	http://www.indecs.com	*DAFS*
Integral Solutions Ltd	http://www.isl.co.uk	*Clementine*
Intrepid Systems	http://www.intrepidsys.com	*DecisionMaster*
Isoft	http://www.alice.fr	*Alice, AC2.*
IVEE Development AB	http://www.ivee.com	*Spotfire Pro*
LEVEL\5	http://www.l5r.com	*Quest*
Logica	http://www.logica.com/products/product-Discovery.html	*Discovery 3D Toolkit*
Magnify	http://www.magnify.com	*PATTERN*
Mathsoft	http://www.mathsoft.com/splus	*S-Plus*
MIT - Management Intelligenter Technologien GmbH	http://www.mitgmbh.de	*DataEngine*
NCR	http://www.ncrhitc.com/high/skills/kd	*Management Discovery Tool (MDT).*
NeoVista Solutions, Inc	http://www.neovista.com	*Decision Series*
NeuralWare, Inc.	http://www.norsys.com	*NeuCOP, NeuralWorks Predict*
Norsys Software Corp	http://www.pilotsw.com	*Netica*
Pilot Software	http://www.prevision.com	*Discovery Server*
Prevision	http://www.quadstone.co.uk	*Strategist, Preclass*
Recognition Systems	http://ourworld.compuserve.com/homepages/reduct	*Relationship Manager*
REDUCT & Lobbe Technologies	http://www.salford-systems.com	*DATALOGIC/R*
Salford Systems	http://www.sas.com/feature/4qdm/intro.html	*CART*

Compañía	URL	Software
Sentient Machine Research B.V.	http://www.snat.de/nc5	*DataDetective*
Siemens Nixdor	http://www.sgi.com/Products/software/MineSet/	*SENN*
Silicon Graphics, Inc.	http://www.slp-infoware.com	*MineSet*
Slp InfoWare	http://www.chemnitz-info.com/softwaregbr/	*User Information System, Customer*
Scholz & Theß Software GbR	http://www.spss.com	*STARC*
Syllogic	http://www.syllogic.nl	*Syllogic Datamining Tool/MP*
TDS Inform	http://www.think.com	*Syllogic Datamining Tool/MP*
Thinking Machines Corporation	http://www.trajecta.com	*Darwin*
Trajecta	http://www.treeage.com	*dbProphet*
TreeAge	http://www.triada.com	*DATA*
Triada	http://www.hal-pc.org/~jpbrown	*Ngram*
Ultimate Resources	http://www.ultranet.com/~unica	*SuperInduction*
Unica	http://www.contact.co.uk/whitecross/white1.htm	*Model 1*
WhiteCross Data Exploration	http://www.wizsoft.com	*HeatSeeker*

En el Anexo 3 se presenta información detallada de los proveedores y el software minero.

2.2 Investigación actualmente desarrollada

Institución:	Instituto de Investigaciones Eléctricas
Fecha de actualización:	07/12/96
Sede de la investigación:	Cuernavaca, Morelos; IIE; Departamento de Sistemas de Información.
Nombre de la investigación:	Estudio y comparación de sistemas de clasificación y de minería de datos.
Objetivo:	Estudiar, analizar y comparar distintos algoritmos y técnicas de clasificación y de minería de datos con la finalidad de identificar aquellos que *efectúen* la clasificación y descubrimientos con mayor eficiencia.
Año inicio:	1996
Año fin:	1996
Grado de desarrollo:	En proceso
Disciplina:	Informática
Responsable:	Garza Melendez, Ricardo
Grado:	Doctorado (Univ. Of Sussex)
Productos de la investigación y descripción (si la hay): 1. Estudio y comparación de distintas técnicas y herramientas para clasificación y minería de datos. 2. Implantación de un sistema prototipo de clasificación de datos y de minería de datos. 3. Documento de una metodología para el diseño de sistemas clasificadores y de minería de datos.	

Institución:	Instituto Politécnico Nacional
Fecha de actualización:	1997
Sede de la investigación:	SoftwarePro International, Austin, Texas.
Nombre de la investigación:	Estado del arte y de la práctica en minería de datos, análisis y crítica.
Objetivo:	Un sistema de minería de datos busca situaciones interesantes, desviaciones, tendencias y anomalías, en un mar de datos. Por lo general la búsqueda es automática (la máquina efectúa los hallazgos sin intervención humana) sobre datos numéricos que yacen en una base de datos relacional .
Año inicio:	1997
Grado de desarrollo:	En proceso
Disciplina:	Informática
Responsable:	Guzmán Arenas, Adolfo
Grado:	Doctorado

Productos de la investigación y descripción (si la hay):

1. Taxonomía de sistemas mineros.
2. Técnicas para manejar grandes volúmenes, proceso incremental.
3. Estructura de una minero: MineDatos.

Estructura de un minero: MineDatos.Esta sección describe cómo esta formado y cómo funciona un MineDatos, un minero diseñado y construido por el autor.Las partes del minero son las siguientes:

1) Selección de variables. Las variables más importantes se escogen de acuerdo al especialista, y se incluyen en el cubo de datos.

2) Cubo de datos. Se linearizan los árboles de cada variable del punto 1. Por ejemplo la jerarquía de fechas (año, mes, semana, día) se coloca sobre un eje. Lo mismo para la jerarquía geográfica, para la jerarquía de productos, o de enfermedades, etc.. Es un hipercubo, pues en general tiene más de tres dimensiones. En cada celda se coloca el totalizador respectivo. Solo los totalizadores más pormenorizados tienen datos inicialmente (es decir, existe un dato de venta para zapatos Bostonianos cafés talla 7 en Salina Cruz el 18 de septiembre de 1997, pero no para esa semana, ni para Oaxaca). Cada celda tiene padres geográficos, de producto, temporales, etc. El cubo de MineDatos es real, es decir, todas sus celdas existen. Primero, existen vacías en su mayoría, y después, son llenadas por adelantado. En una segunda versión, las celdas son llenadas bajo demanda (solo si lo necesitan).

3) Se definen los mineros como clientes del servidor de datos, que está en Infromix. Los clientes son programas en C que evalúan una expresión (predicado) que regresa V ó F sobre cada celda. Los mineros efectúan un barrido sobre las celdas; aquellas que producen V originarán una "situación interesante" que será registrada en un archivo de salida. La expresión a evaluar está en forma normal posfija, y no se convierte a SQL, sino que se interpreta.

Es decir, el minero analiza su predicado y extrae del cubo real un sub-cubo de datos, trayéndolo a la memoria de la máquina cliente para su procesado. La mayor parte del tiempo los mineros están *haciendo sumas para calcular los totalizadores* que no existen en el cubo. Esta arquitectura aprovecha el poderío de las varias máquinas conectadas al servidor; éste se utiliza sólo para repartir bloques de datos a las máquinas que son las que hacen el trabajo pesado. (*Muchos* mineros comerciales efectúan este trabajo en el servidor, sobrecargándolo).

4) Cada minero tiene una agenda de trabajos pendientes (regiones a escudriñar). Cuando los trabajos prioritarios aumentan, el minero cesa de trabajar, reanudando sus labores posteriormente.

5) Se les puede cambiar a los mineros:

a) Las fórmulas de los predicados. Inicialmente, buscan por máximos, mínimos, pendientes positivas o negativas "considerables", y otras cuantas "curvas" interesantes. En este sentido, *los mineros hacen ajuste de curvas.*
b) La región de búsqueda. Por ejemplo, restringiendo solo la zona sur del país.
c) EL orden de búsqueda. Por ejemplo, mirar primero tal cual producto o enfermedad. d) Interacción diferida con el usuario. Si el usuario manualmente nota que una celda esta cerca de ser interesante, puede fabricar un minero ad hoc que la vigilará, avisándole cuando sea interesante.

6) La interacción que un sistema de mineros brinde al usuario debe ser lo más amigable posible, pues él *normalmente no tiene los conocimientos* requeridos de estadística, informática, de cómo está el cubo, etc.

El campo de la minería de datos es nuevo ya ha despegado al abaratamiento de la capacidad de cómputo de la computadoras personales y de la de almacenamiento del disco de cabeza móvil. En general, es un campo donde el estado de la práctica (*programas y sistemas comerciales*) está dominado por consideraciones de eficiencia y utilidad, en tanto que el estado del arte (publicaciones e investigación) esta desvinculado de los practicantes, por no saber quiénes son. Por consiguiente, es un *campo muy fructífero para instituciones* o grupos donde ambos aspectos se combinen o complementen.

Algunos desarrollos ya conocidos, como clasificadores, visualización, e incluso selección de variables, son ofrecidos como parte de un sistema de minería de datos, aunque nosotros reservamos este nombre para la búsqueda automática de *situaciones interesantes, desviaciones* y anomalías.

Las *aplicaciones que se procesan en línea* (OLAP en inglés) no son necesariamente mineros, aunque algunos mineros son aplicaciones en línea.

2.3 Bibliografía en venta

USAMA M. Fayyad, Ramasamy **Uthurusamy** (Editor), Gregory **Piatetsky-shapiro** (Editor), Padhraic **Smyth** (Editor), Advances in Knowledge Discovery & Data, Editorial M I T Press, Marzo1996, ISBN 0262560976

PALAZ, Ibrahim, Salles **Sengupta,** Automated Pattern Analysis in Petroleum Exploration, Editorial Springer-Verlag New York, Incorporated, Enero1992, ISBN 0387974687

PETER Adriaans, Dolf **Zantinge,** Data Mining, Editorial Addison-We, Septiembre 1996, ISBN 0201403803

GROTH Robert, Data Mining: A Hands on Approach to Information Discovery, Editorial Prentice Hall, Agosto1997, ISBN 0137564120

BERRY Michael, Gordon **Linoff,** Data Mining Techniques: For Marketing, Sales & Customer Support, Editorial John Wiley & Sons, Incorporated, Jun1997, ISBN 0471179809

P. BIGUS Joseph, Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support, Editorial McGraw-Hill Companies, Jun1996, ISBN 0070057796

M. MATTISON Robert, Data Warehousing and Data Mining for Telecommunications, Editorial Artech House, Incorporated, Agosto1997, ISBN 0890069522

CABENA Peter, Pablo **Hadjnian,** Discovering Data Mining from Concept to Implementation, Editorial Prentice Hall, Agosto1997, ISBN 0137439806

H-J Lu, H. Motoda, H. **Liu,** Knowledge Discovery & Data Mining: Techniques & Applications, Editorial World Scie, Mayo1997, 9810230729

M. WEISS Sholom, Nitin **Indurkhya,** Predictive Data Mining, Editorial Morgan Kau, Agosto1997, ISBN 1558604030

Institute Series F Inc. Unica Technologies, Solving Data Mining Problems Using Pattern Recognition Software (Data Warehousing), Editorial Prentice Hall, Diciembre 1997, ISBN 0130950831

J. Komorowski, Jan M. **Zytkow,** Principles of Data Mining & Knowledge Discovery: Proceedings of the First European Symposium, Pkdd '97, Trondheim, Norway, June24-27, 1997, Vol. 126, Editorial
Springer-Verlag New York, Incorporated, Octubre 1997, ISBN 3540632239

USAMA M. Fayyad, Ramasamy **Uthurusamy** (Editor), Proceedings of the Third International Conference on Knowledge Discovery & Data Mining, Editorial AAAI Pr, Enero1995, 0929280822

T. Y. Lin (Editor), Rough Sets & Data Mining: Analysis of Imprecise Data, Editorial Kluwer Academic Publishers, Noviembre 1996, ISBN 0792398076

FAYYAD, Usama M. and Piatetsky-**Shapiro,** Gregory and Smyth, Padhraic and Uthurusamy,Ramasamy, Advances in Knowledge Discovery and Data Mining, Editorial MIT Press, 1995

PIETER Adriaans, and Dolf **Zantiage**, Data Mining, Editorial Addison-Wesley, 1996

BIGUS, Joseph P, Data Mining with Neural Networks, Editorial McGraw-Hill, 1996

HAMMERGREN, Thomas C., Data Warehousing: Building the Corporate Knowledgebase, Editorial International Thomson Computer Press, 1996

2.4 Conclusiones

Las fuentes de información consultadas fueron de gran utilidad para el desarrollo del presente trabajo de investigación. La más consultada fue la Internet en la que obtuvimos más del 50% de la información que se presenta.

Desafortunadamente no nos fue posible obtener libros de Data Mining en el D.F.. Unicamente pudimos conseguir uno traído de los Estados Unidos.

La minería de datos es un concepto que se esta difundiendo mucho entre todos los ámbitos empresariales de hoy en día, lo cual pudimos observar en de la gran cantidad de conferencias y seminarios, del tema y relacionados con el mismo, que se han dado a la fecha. Sin embargo, la mayoría de dichos seminarios y conferencias han sido organizadas y llevadas a cabo por empresas extranjeras.

En nuestra búsqueda por la Internet encontramos muchos proveedores de software minero de datos.

Pocas son las empresas, en México que se dedican al desarrollo de software minero. Así como pocas son las que lo usan y al parecer, aún esta en el ámbito académico y de investigación.

Marco conceptual

3. MARCO CONCEPTUAL

3.1 Antecedentes

La humanidad se ha interesado por los datos, al menos durante los pasados 12 000 años, y aunque en la actualidad a menudo se asocia el concepto de datos con la computadora; históricamente han existido otros métodos primitivos de manejo de datos, en realidad, algunos todavía siguen utilizándose.

En retrospectiva, los orígenes primarios del interés en datos puede seguirse hasta el surgimiento de las ciudades. Los principios de la producción en masa, la especialización de la mano de obra, el empleo del dinero, y la posibilidad de alquilar servicios y productos para las necesidades de vida requerían la conservación de datos en registros. Conforme pasó el tiempo, se conservaron más tipos de datos y registros diferentes; estos incluían calendarios, datos censales, investigaciones, registros de propiedad de tierra y de matrimonio, datos acerca de contribuciones a la iglesia, y árboles genealógicos.

El nivel de complejidad comercial fue aumentando, lo que a su vez provocó otro aumento de conservación de registros cada vez más complejos: específicamente la teneduría de libros de partida doble. Fue también en el siglo XVII que las personas empezaron a interesarse en dispositivos que pudieran procesar sus datos automáticamente, aunque fuera en forma rudimentaria.

El desarrollo de dispositivos de cómputo y tabulación prácticos requirió de dos cosas: 1) una capacidad de almacenamiento y procesamiento de datos a gran escala y 2) un cierto nivel básico de capacidades de manufactura y dispositivos mecánicos y eléctricos. Hacia fines del siglo XIX, las capacidades de maquinado habían llegado a un nivel de avance suficiente y ocurrió también una necesidad que comprendía una cantidad masiva de datos.

Paralelo al crecimiento del equipo para procesar datos se dio el desarrollo de nuevos medios para almacenarlos. La primera forma moderna de almacenamiento de datos fueron las perforaciones en algún medio adecuado de papel. En las décadas de 1870 y 1880 la cinta de papel perforada se presentó junto con los primeros equipos de teletipo. Hollerith hacia 1890 y Powers a principios de este siglo utilizaron tarjetas perforadas como medio de almacenamiento.

Hacia 1936 tanto Eastman Kodak como Bush, con su Rapid Selector, registraban datos digitales en forma fotográfica y aunque estos podían leerse de nuevo con tecnología de celdas fotoeléctricas, éste era un medio no borrable.

En la segunda mitad de la década de 1930 se inició la era de los medios de almacenamiento magnéticos con los experimentos Bell Laboratories con cinta magnética para almacenamiento de sonido; en 1937 se realizaron trabajos, en Alemania, con óxidos para utilizarlos en el almacenamiento magnético.

Específicamente en 1947, Eckert y Mauchly desarrollaron una unidad de cinta magnética. En 1952, tanto la UNIVAC como Raytheon ofrecieron en el mercado en forma comercial unidades de cinta magnética.

La era actual se caracteriza por generar, recopilar y procesar información. La cantidad de estas actividades que hoy en día se realizan no tiene precedentes en la historia. Más aún, el volumen aumenta a un ritmo vertiginoso y continuará así en el futuro. La tecnología de la computación electrónica ha dejado impresa su huella en el mundo y sus avances continuarán demarcándola.

Las organizaciones públicas privadas y los ciudadanos de muchas naciones se ven afectados y beneficiados cada vez más por la tecnología computarizada que almacena, maneja y procesa datos. Paulatinamente se aprecian más en forma directa, los beneficios de esa tecnología. Una gran parte de esa información que se recopila no se computariza aún, o si se computariza no se explota. El costo del equipo y de los programas para su manejo, decrece a un ritmo muy rápido, al tiempo que la cantidad de información que almacenan y administran las computadoras crece desmesuradamente. Puede verse que ya una vasta porción de los requerimientos actuales de datos se almacenarán más económicamente en bancos de datos y bases de datos de computadoras que en papel.

No sólo se almacena información impresa o tabular; esta tecnología en desarrollo maneja cada vez más: dibujos con líneas, fotografías, grabaciones de voz y otros tipos de información.

En 1968, Ascher Olper, que en ese momento pertenecía al personal de IBM, advirtió que era necesario unificar el conocimiento sobre la programación de manera que estuviese al alcance de todos los programadores de sistemas.

En respuesta a la petición hecha por de Olper IBM, decide patrocinar la serie sobre programación de sistemas, un ambicioso proyecto cuya finalidad era reunir, organizar y publicar los principios y las técnicas que tuvieran utilidad permanente en la industria en general. [1]

El hecho de asociar una estructura de datos con una relación fue identificado por investigadores y documentados en el año de 1967, por R.E. Levein y M.E. Maron. Posteriormente, D.L. Childs, en 1968, efectuó una siguiente publicación al respecto. Pero es hasta después del artículo de junio 1970 en el volumen 13, número 6 de Comunications of the ACM donde aparece el articulo "A Relational Model of Data for Large Shared Data Banks" de Eddgar F. Codd que da por iniciado el modelo de datos relacional.

Ese artículo fue el resultado de un trabajo que se inició en 1968 en el IBM Reserach Laboratory en San José, California, con un modelo abstracto de información. El objetivo era encontrar un fundamento teórico de los diferentes aspectos de un Data Base Management System completamente ajeno de los aspectos de un proceso físico de una máquina o CPU en particular. Este modelo es el que actualmente se conoce como modelo relacional o estructuras de datos relacionales.

En 1969, el Dr. E. F. Codd publicó el primer documento que define un modelo de bases de datos basado sobre el concepto matemático de los conjuntos relacionales. El Modelo de Bases de Datos Relacional (RDBM) ha sido refinado constantemente desde entonces, y mucho más a partir de la publicación que hizo Codd de las "12 reglas" para las bases de datos relacionales en 1985, y en 1990 su libro que define la versión 2 (RV/2) del modelo relacional mediante 333 reglas que son un subconjunto y expansiones de las 12 originales.

A mediados de los años 70´s IBM desarrolló como lenguaje estándar el Lenguaje de Consulta Estructurado (SQL), fue diseñado como un lenguaje para acceder explícitamente a DBMS´s basados en el modelo relacional que corría sobre mainframes. La versión inicial del lenguaje apareció como SEQUEL, posteriormente su nombre fue reducido a SQL. [2]

3.2 Definiciones

3.2.1 Etimológica

Término	Definición
Mina Mena	Del latín minare
Menare:	Conducir
Mina:	Conducto: conducto para buscar minerales
Del griego Myo:	Oculto
	Mina: lo oculto.
Derivados:	Minador, minal, minar, minera, mineraje, mineral, minería, minero.
Eria:	Desinencia general
Minería:	Lo relativo al oficio.
Ero:	Denota en los sustantivos (mina), la profesión, oficio, ocupación.

3.2.2 De diccionario

Minería: Excavación que hace por pozos, galerías y socavaciones a cielo abierto para extraer mineral.

Aquello que abunda en cosas dignas de aprecio o de que puede sacarse algún provecho o utilidad.

Oficio, empleo o negocio de que con poco trabajo se saca mucho interés y ganancia.

Datos: Del latín Datum, lo que se da. Antecedente necesario para llegar al conocimiento exacto de una cosa o para deducir las consecuencias legítimas de un hecho.

Noticia o antecedente necesario para formar un concepto de algún asunto y resolverlo con acierto. [3]

3.2.3 De autores

Para William J. Frawley, Gregory Piateski-Shapiro y Christopher J. Matheus:

"Minería de datos o recuperación del conocimiento en bancos de datos como también se le conoce, es la extracción implícita no trivial previamente conocida y potencialmente útil de los datos. Esto comprende un número diferente de aproximaciones técnicas diferentes como índices, resumen de la información, reglas de clasificación del aprendizaje, dependencia de redes, análisis de cambios y detección de anomalías". [4]

Para Macel Holshemire y Arno Siebes (1994):

"Minería de datos es la búsqueda de relaciones y patrones globales que existen en grandes bases de datos pero son ocultos entre gran cantidad de datos, tal como una relación entre datos del paciente y el diagnóstico de su médico. Estas relaciones representan valioso conocimiento acerca del banco de datos y los objetos en el banco de datos". [4]

Para Adolfo Guzmán Arenas:

"Un sistema de minería de datos está formado por varios programas de cómputo, que realizan la búsqueda en una base de datos, de manera automática, de tendencias, desviaciones, anomalías, patrones y situaciones "interesantes". Estas desviaciones o tendencias a menudo son reportadas inmediatamente, o más a menudo en un archivo, para su posterior visualización y decisión final. Existen varios algoritmos generales para "minería de datos". El sistema a menudo posee un configurador que permite particularizarlos a casos y situaciones específicas; una agenda de trabajo guía a cada minero en sus búsquedas. La prioridad de búsqueda (orden en la agenda), los criterios de "interés" y "tolerancia" y las definiciones de "situación anómala" son modificables por el usuario, de suerte que un minero originalmente posee un criterio general pero quizá fuera de foco de lo que es "interesante" para el usuario, y termina como una colección de límites de valores y criterios de éxito muy específicos, que busca en determinada área de la base de datos y bajo condiciones o predicados igualmente ajustados a la medida ". [5]

Para Harjinder S. Gill y Prakash C. Rao:

"La minería de datos apoya la modalidad de descubrimiento del soporte de decisiones. Las herramientas de minería de datos recorren los datos detallados de transacciones para desenterrar patrones y asociaciones ocultos. Por lo regular los resultados generan extensos reportes o se les analiza con herramientas de visualización de datos descubiertos". [4]

3.2.4 Propia

Para Ma. Isabel Angeles Larrieta y Angélica Ma. Santillán Gómez:

Minería de datos es el proceso cuyo objetivo es descubrir, extraer y almacenar información relevante y previamente ignorada de amplias bases de datos a través de programas de búsqueda e identificación de patrones y relaciones globales, tendencias, desviaciones y situaciones interesantes que se encuentran ocultas en grandes cantidades de datos y que pueden ser descubiertas mediante diferentes técnicas de análisis, tales como: cluster, asociaciones, clasificación, visualización, redes neuronales, algoritmos genéticos, detección de desviaciones, entre otros.

La idea es aprovechar el valor de la información encontrada y usar los patrones de ésta para que el tomador de decisiones obtenga un mejor conocimiento de su negocio.

3.3 Sinónimos

Palabra	Sinónimo[6]
Mina	Yacimiento
	Filón
	Galería
	Almacén
	Excavación
	Túnel
Minar	Socavar
	Excavar
	Horadar
Dato	Nota
	Detalle
	Antecedente
	Noticia
	Documento

3.4. Evolución

La minería de datos es el resultado de un largo proceso de búsqueda. Esta evolución inició con el almacenamiento de datos en computadoras, continuó con el mejoramiento en el acceso de los datos almacenados, y más recientemente con la generación de tecnologías que permita a los usuarios navegar a través de sus datos en tiempo real.

En la evolución de la tecnología de información cada nuevo paso fue construido a partir de uno previo. Lo cual lo podemos observar en el acceso dinámico a los datos que fue un antecedente necesario para la navegación de aplicaciones y en la habilidad para almacenar grandes bases de datos que fue el antecedente de la minería de datos

El siguiente cuadro muestra los cuatro pasos evolutivos de la minería de datos. [7]

Pasos Evolutivos	Cuestionamientos de la empresas	Tecnología disponible
Almacenamiento de los datos (1960s)	¿Cuál fue mi ganancia total en los últimos cinco años?	Computadoras, cintas y discos.
Acceso a los datos (1980s)	¿Cuáles fueron mis ventas totales en la sucursal Centro el mes de marzo?	Bases de datos relacionales (RDBMS), Lenguajes de consulta estructurada (SQL).
Datawarehousing y soporte de decisiones (1990s)	¿Cuál fue la diferencia de ventas en la sucursal Centro en relación a la sucursal Sur en el mes de Marzo?	Procesamiento analítico en línea (OLAP), Bases de datos multidimensionales, Datawarehouse
Minería de datos (Desarrollandose hoy en día)	¿Qué es lo más probable que pase en las ventas de la sucursal Centro en el próximo mes de Junio? ¿Porqué?	Algoritmos avanzados, computadoras multiproceso, bases de datos masivas.

A continuación se hace una descripción de cada uno de los pasos evolutivos de la *minería* de datos.

3.4.1 Almacenamiento de los datos (1960s)

En un principio las empresas tenían como principal preocupación el almacenamiento de sus datos, debido a que generaban grandes cantidades de los mismos era necesario diseñar alguna forma en la que pudieran ser guardados; por lo que se creó el concepto de banco de datos, el cuál es descrito a continuación:

3.4.1.1 Banco de datos

Son un conjunto de registros listados uno después de otro, almacenados en archivos. Al hablar de bancos de datos se ha considerado que estos existen totalmente independientes entre sí. En la realidad, el manejo de combinaciones de tales archivos resulta difícil porque los datos se almacenan en diferentes formatos y en distintos archivos, una de las características de los archivos es que los datos no se pueden compartir entre los diferentes programas, a menudo no son recuperables ni están seguros.

Hay cuatro acciones que pueden realizarse con un archivo: 1) simplemente consultarlo sin modificarlo, 2) Modificar un registro, 3) Insertar un nuevo registro y 4) Borrar un registro existente.

Existen varias formas de organizar y accesar archivos para su recuperación subsecuente.

Organización de archivos:

- ***Archivos secuenciales***: La forma más simple de almacenar un conjunto de registros es en una lista larga. Si los registros están en secuencia, en términos de

uno o más campos, se dice que el arreglo es un archivo secuencial. En un archivo secuencial la única forma de recuperar los datos es empezar al inicio de archivo y leer un registro después de otro hasta llegar al final.

- ***Archivos no secuenciales***: Tienen la ventaja de que no es necesario dedicar tiempo o esfuerzo para mantener el archivo en secuencia física al unir, clasificar o realizar cualquier otra manipulación, ya sea al crear el archivo por primera vez, o cuando más adelante se agregan nuevos registros al final del archivo.

- ***Archivos indizados***: Se espera tener acceso a esos registros en forma directa, se espera recuperar aquellos registros en secuencia de acuerdo con algún campo, ya sea que el archivo esté o no almacenado en secuencia física, de acuerdo con ese u otro campo.

Desafortunadamente la sencillez de los bancos de datos (independientes), deja mucho que desear en cuanto a eficiencia a algunas situaciones comunes del procesamiento y almacenamiento de datos. [1]

3.4.2 Acceso a los datos (1980s)

El concepto de bases de datos surgió a partir del manejo de combinaciones de los bancos de datos, llevando a un concepto más poderoso de almacenamiento de datos.

3.4.2.1 Bases de datos

Una base de datos es una colección de datos, que guardan una relación entre sí. Un sistema de bases de datos, es un sistema de mantenimiento de registros basado en computadoras, es decir, un sistema cuyo propósito general es registrar y mantener información. Un sistema de bases de datos incluye cuatro componentes principales: datos, hardware, software y usuarios.

Los puntos clave que se tomaron en cuenta para la creación de bases de datos fueron: ciertas características del medio ambiente de procesamiento de datos, los problemas asociados para almacenar datos redundantes, la capacidad para almacenar datos que tiene que ver con relaciones múltiples y el concepto de independencia de datos.

Con forme creció el procesamiento de datos empezaron a cambiar un cierto número de reglas básicas. El hardware se volvió más barato, el desarrollo del software tomó una forma más estandarizada y estructurada, y se acumularon muchas aplicaciones nuevas que debían implantarse.

Con el medio ambiente de las bases de datos se logró compartir archivos de datos entre distintas aplicaciones y emplear descripción de datos y proposiciones de acceso consistentes en los programas, obligando a tener una visión más cuidadosa y estandarizada de todo el proceso de desarrollo de la aplicación. [7]

Así, pues, las bases de datos proporcionan el marco para tratar los datos como un recurso estandarizado, administrable y compartible, preocupándose por la seguridad, respaldo y recuperación, concurrencia y capacidad de auditoría de los datos.

3.4.2.2 DBMS (Data Base Management System)

Un sistema manejador de bases de datos (DBMS) es probablemente mejor definido como una pieza sofisticada de software, la cual soporta la creación, manipulación y administración de sistemas de bases de datos. Un sistema de bases de datos constituye en sí mismo un sistema completo de información.

Los elementos de un DBMS son:
1) Lenguaje de definición de datos (DDL - Data Definition lenguage).
2) Diccionario de Datos (DD - Data Dictionary).
3) Lenguaje de manipulación de datos (DMl - Data Manipulation Lenguage).

Un sistema de bases de datos consta de dos partes: el **Sistema Manejador de Bases de Datos (DBMS)**, el cual es el programa que organiza y mantiene estas listas de información, y la **Aplicación de Bases de Datos**, la cual es un programa que nos permite recuperar, consultar y modificar la información almacenada por el DBMS.

Para llevar a cabo el almacenamiento de datos, el DBMS cuenta con algún tipo de servicio de definición de datos para definir los registros y campos en la base de datos. Además se necesita un mecanismo interno que mantenga los datos en el disco y conozca en dónde reside cada elemento. Un DBMS provee los siguientes servicios:

- ***Definición de datos:*** Provee un método de definición y almacenamiento de una cierta cantidad de datos.
- ***Mantenimiento de datos:*** Mantiene los datos utilizando un registro para cada elemento, los campos contienen información particular que describe cada elemento.
- ***Manipulación de datos:*** Provee servicios que permiten al usuario insertar, modificar, borrar y ordenar los datos de la base de datos.
- ***Integridad de datos:*** Provee uno o más métodos para asegurarse que los datos son correctos.
- ***Despliegue de datos:*** Provee algún método para presentarle los datos al usuario.

3.4.2.3 Modelos de bases de datos

Existen cuatro tipos principales de modelos de bases de datos para definir registros o entidades y las relaciones que guardan, esto es, las estructuras de las bases de datos para visualizar y manipular los datos a nivel lógico:

- Sistema Manejador de Archivos (**FMS** File Management System)
- Sistema de Base de Datos Jerárquico (Hierarchical Database System)
- Sistema de Base de Datos de Red (Network Database System)
- Modelo de Base de Datos Relacional (**RDBM** Relational Database Model)

- Un nuevo tipo de DBMS es el llamado Orientado a Objetos (Object Oriented Databases), el cual ha sido muy difundido en los últimos dos años.

3.4.2.3.1 Sistema Manejador de Archivos (FMS)

Es el modelo de base de datos más sencillo de comprender y es el único que describe como los datos están almacenados en el disco. En este modelo, cada campo o dato es almacenado secuencialmente sobre el disco en un archivo muy largo. El sistema manejador de archivos fue el primer método utilizado para almacenar datos en una base de datos computarizada. Las desventajas de este modelo son claras; primero no hay indicación de la relación entre los elementos más que la secuencia de almacenamiento. El programador y a veces el usuario tiene que conocer exactamente como los datos están almacenados en el archivo y el orden en el que pueden ser accesados. La segunda desventaja del FMS es que crea problemas con la integridad de los datos; el valor de todos los archivos tienen que ser verificados por el programa de aplicación antes de ser almacenados en el disco, además que no hay forma de buscar un registro específico rápidamente, cada búsqueda debe comenzar desde el principio del archivo examinando registro por registro. La desventaja más grandes del FMS es que no permiten cambiar fácilmente la estructura de la base de datos.

3.4.2.3.2 Modelo de Base de Datos Jerárquica

En este modelo los datos son organizados en una estructura de árbol que se origina desde una raíz. Cada clase de datos es localizada en diferentes niveles de una rama particular que proviene de la raíz. La estructura de datos en cada nivel de clase es llamado *nodo*; si no nacen de él nuevas ramas el último nodo en las serie es considerado una *hoja*.

En términos de bases de datos, la estructura de árbol define las relaciones padre - hijo y hermano entre los distintos elementos en la base de datos y esto muestra una ventaja sobre los sistemas manejadores de archivos (FMS), ya que permiten definir las relaciones uno a muchos. Además la estructura jerárquica hace fácil y rápida la búsqueda de datos.

En una estructura jerárquica siempre existe un nodo raíz único el cual es generalmente propietario por el sistema o el DBMS. Los nodos de nivel 1 representan una clase particular de datos. Cada nodo de nivel 1 puede tener uno o más hijos de nivel 2.

La estructura física de los datos en el disco no tiene importancia en el modelo jerárquico; el DBMS puede almacenar los datos como una lista ligada de campos, con apuntadores que van del padre al hijo y de rama a rama, finalizando en un valor nulo o apuntador terminal en la última hoja.

Este diseño facilita la adición de nuevos nodos en cualquier nivel, ya que el DBMS únicamente tiene que modificar el apuntador terminal al siguiente nodo de la rama en la lista. Por conveniencia, podemos definir un registro como un padre y todos sus hijos.

El primer problema radica en la estructura inicial de la base de datos, la cual es arbitraria y debe ser definida por el programador cuando la base de datos es creada. La *relación padre-hijo no puede ser modificada* sin rediseñar la estructura. Otro problema creado por la rigidez de la estructura jerárquica, es la dificultad para modificar la definición de los niveles de clases, ya que se tiene que redefinir la estructura.

La desventaja más significativa de este modelo es que no provee un *método de* definición sencillo para el uso de relaciones muchos a muchos; una solución a este problema es el almacenamiento de múltiples copias del mismo dato en múltiples niveles. Otro enfoque de solución al problema de la relación muchos a muchos es ir aumentando relaciones padre-hijo secundarias y apuntadores en la estructura jerárquica.

3.4.2.3.3 Modelo de Base de Datos de Red

Las primeras *especificaciones fueron escritas en 1971 por* CODASYL. Hay que notar, que el término "red" no tiene relación con el medio físico en el que actualmente corren las bases de datos, el modelo de red define conceptualmente las bases de datos en las cuales existen las relaciones muchos a muchos. Las relaciones entre los diferentes datos son referidas comúnmente como conjuntos que los distinguen estrictamente de las relaciones padre-hijo definidas en el modelo jerárquico.

Un modelo de base de datos de red se identifica por líneas o apuntadores cíclicos para mapear las relaciones entre los diferentes elementos de datos. El modelo de red puede ser *utilizado* para describir relaciones cada vez más complejas. La flexibilidad del modelo de red en las relaciones muchos a muchos es fuerte.

Las interrelaciones entre los diferentes conjuntos pueden convertirse en un modelo cada vez más complejo y difícil de *mapear. Tal como las bases de datos jerárquicas, las* bases de datos de red pueden ser muy rápidas, especialmente mediante el uso de índices de apuntadores que permiten la ubicación directa en el primer elemento de un conjunto en una búsqueda.

Sin embargo, el modelo de red sufre del mismo problema estructural mencionado en la descripción del modelo de sistemas jerárquico. El diseño inicial de la base de datos es arbitrario, una vez que este es instalado cualquier cambio requiere crear una nueva estructura. El modelo de red permite adicionar nuevos datos o modificaciones a los ya *existentes de manera simple*, ya que sólo se tiene que definir un nuevo conjunto de relaciones propias con el resto del conjunto de datos.

3.4.2.3.4 Modelo de Base de Datos Relacional

El modelo relacional abandona el concepto de relaciones padre-hijo entre diferentes elementos de datos. Además, el dato es organizado en conjuntos lógicos matemáticos dentro de una estructura tabular. En un RDBM, cada campo se convierte en una columna dentro de una tabla, y cada registro se convierte en un renglón. Diferentes relaciones entre varias tablas, son definidas a través del uso de funciones matemáticas, tales como el *JOIN* y *UNION*.

Cada tabla tiene una o más columnas con el mismo nombre que se encuentra en otra tabla; son estos nombres de columnas comunes los que son utilizados para relacionar diferentes tablas. Sin embargo, los nombres de columnas no tienen que ser idénticos en el modelo relacional.

El modelo relacional tiene distintas ventajas sobre el modelo jerárquico y de red, la más importante de las cuales es su completa flexibilidad en la descripción de las relaciones entre los diferentes elementos de datos. El programador define la base de datos creando las tablas y decidiendo cuáles columnas serán relacionadas. Desde este punto, los usuarios pueden consultar la base de datos sobre alguna de las columnas en una tabla o sobre las relaciones entre diferentes tablas. Modificar la estructura de la base de datos es tan simple como aumentar o borrar columnas de una tabla, lo cual no afecta a otra tabla de ningún otro modo. Pueden ser creadas nuevas tablas como proyecciones (subconjuntos) de tablas existentes, y otras tablas pueden ser removidas.

No se tiene que reconstruir la estructura de la base de datos completa para hacer cambios, esto representa un incremento en la preservación de la integridad de datos.

3.4.2.4 Arquitecturas de DBMS

El tipo de sistemas de computadoras en que las bases de datos pueden ejecutarse, pueden ser clasificados en cuatro categorías o plataformas principales: centralizada, PC, Cliente/Servidor y distribuidas. La arquitectura del DBMS mismo no determina necesariamente el tipo de sistema computacional en que la base de datos tiene que ejecutarse.

3.4.2.4.1 Plataforma Centralizada

En un sistema centralizado, todos los programas se ejecutan sobre una computadora principal, incluyendo el DBMS, las aplicaciones que accesan a la base de datos y las facilidades de comunicación que envían y reciben datos de las terminales de usuarios.

Los usuarios accesan la base de datos a través de una conexión local o terminales remotas. Las terminales son generalmente mudas, tienen o no poder de procesamiento propio, y constan únicamente de una pantalla, teclado y hardware para comunicarse con el host. La ventaja de los microprocesadores ha llevado al desarrollo de terminales más inteligentes en los últimos años, donde la terminal comparte responsabilidad para manejar la salida en pantalla y la entrada de datos por parte del usuario.

Mientras los sistemas en mainframe y microcomputadoras son la plataforma principal para sistemas de bases de datos corporativas, los sistemas basados en PC´s pueden comunicarse además con sistemas centralizados a través de comunicaciones de hardware/software que emulan (imitan) los tipos de terminal utilizados con un host particular.

Todo el procesamiento de datos en un sistema centralizado toma lugar en el host, y el DBMS debe estar corriendo antes de que las aplicaciones puedan accesar la base de datos.

Las principales ventajas de un sistema centralizado es su seguridad y la habilidad de manejar enormes montos de datos en dispositivos de almacenamiento. Además, soportan un gran número de usuarios simultáneos. Las desventajas se refieren generalmente a los costos de instalación y mantenimiento. Grandes sistemas de mainframes y de minicomputadoras requieren soporte especializado.

3.4.2.4.2 *Sistemas de Bases de Datos Cliente/Servidor*

De manera sencilla, una base de datos Cliente/Servidor divide el procesamiento de la base de datos entre dos sistemas: el cliente (el cual ejecuta la aplicación de la base de datos) y el servidor (el cual ejecuta todo o parte del DBMS actual). El servidor de archivos LAN provee recursos compartidos, tales como espacio en disco para las aplicaciones, e impresoras. El servidor de bases de datos puede encontrarse corriendo en la misma PC como el servidor de archivos.

La aplicación del cliente, se identifica como *front-end*, maneja todas las pantallas y el procesamiento de entrada y salida del usuario. El *back-end* del servidor de bases de datos maneja el procesamiento de datos y el acceso a disco.

Por ejemplo, un usuario en el front-end crea un requerimiento de datos (consulta) para el servidor, y la aplicación del front-end envía el requerimiento a través de la red al servidor. El servidor ejecuta la búsqueda y envía de regreso únicamente los datos que corresponden a la solicitud del usuario.

La ventaja de un sistema cliente/servidor resulta obvia, la división del procesamiento entre dos sistemas reduce la cantidad de tráfico de datos en la red. Mientras los sistemas cliente corren generalmente sobre PC´s, el servidor de bases de datos puede correr sobre otra PC o mainframe.

La desventaja de los sistemas de bases de datos descritos anteriormente es que requieren que los datos se encuentren almacenados en un sistema único. Esto puede ser un problema cuando se tiene la necesidad de atender a usuarios dispersos en un área geográfica, o bien, que necesitan compartir porciones de sus bases de datos departamentales con otros departamentos o un host central.

3.4.2.4.2.1 La tecnología Cliente/Servidor

Un sistema cliente/servidor es aquel que divide el procesamiento de datos entre dos componentes distintos. Por esta definición, los sistemas cliente/servidor no se limitan a aplicaciones de bases de datos; una aplicación que tiene una interfaz de usuario (front-end) y que corre localmente sobre un cliente y una parte del procesamiento corre sobre el servidor (back-end) es una forma de cómputo cliente/servidor.

Actualmente el enfoque principal es sobre aplicaciones de bases de datos. La razón de ello es que una gran cantidad de empresas están reduciendo sus costos de cómputo mediante downsizing de sus bases de datos. La idea del downsizing es simple, mover las bases de datos corporativas de sistemas grandes y centralizados, a sistemas pequeños y menos costosos que no requieren de soporte y mantenimiento extensivos. La división en el poder de procesamiento, que es la base de la arquitectura cliente/servidor, hace posible que los sistemas pequeños manipulen los datos.

Las principales ventajas de un sistema cliente/servidor surgen de la división del procesamiento entre el sistema cliente y el servidor de bases de datos. Desde que el procesamiento de la base de datos se lleva a cabo en el back-end, la velocidad del DBMS no se encuentra ligada con la velocidad de la estación de trabajo. Como resultado, la estación de trabajo necesita estar habilitada para correr el software front-end.

Otro beneficio que se obtiene, es la independencia entre estaciones de trabajo; los usuarios no se encuentran limitados a un solo tipo de sistema o plataforma. En un sistema cliente/servidor, las estaciones de trabajo pueden ser PC´s compatibles con IBM, Macintosh, estaciones de trabajo UNIX, o una combinación de las anteriores, y pueden correr múltiples sistemas operativos, tales como MS/PC-DOS, MS Windows, IBM OS/2, o System 7 de Apple.

Otra ventaja es la preservación de la integridad de datos. Actualmente, varios servidores de bases de datos corren en un DBMS con base en el modelo relacional, de modo que los usuarios no pueden accesar a los datos fuera del DBMS. Además, el DBMS puede proveer servicios de protección de datos, tales como almacenamiento de archivos encriptados; respaldos en cinta en tiempo real, lo cual ocurre mientras la base de datos está siendo accesada; generación de discos espejo, en el cual los datos son escritos automáticamente en una base de datos duplicada sobre un disco duro distinto.

Por otra parte, el DBMS puede proveer procesamiento de transacciones, lo cual permite tener un registro de los cambios a la base de datos y ayuda a corregir los errores que se presenten en la base de datos en caso de que el servidor tenga algún problema.

El procesamiento de transacciones es un método por el cual el DBMS mantiene un log (registro) de todas las modificaciones hechas a la base de datos en un periodo determinado. Se utiliza principalmente para las bases de datos que están siendo modificadas constantemente, tal como un sistema de procesamiento de órdenes, para asegurarse que las modificaciones a los datos están siendo almacenadas adecuadamente en la base de datos.

El log es utilizado para restaurar la base de datos (tanto como sea posible) debido a un estado de error por el cual el sistema haya fallado durante alguna modificación. El DBMS es responsable de manejar la seguridad necesaria para prevenir múltiples cambios al mismo registro o campo; los conflictos y deadlocks entre usuarios que modifican el mismo registro son reducidos significativamente cuando son manejados por un DBMS central.

3.4.2.4.3 Sistemas de Procesamiento Distribuido

El procesamiento distribuido ha existido en su forma simple desde hace varios años; de una manera muy limitada, los datos son compartidos entre varios hosts mediante el envío a través de conexiones directas (sobre la misma red) o a través de conexiones remotas vía telefónica o líneas de datos dedicadas. Una aplicación que es ejecutada sobre uno o más hosts extrae la porción de datos que ha sido modificada durante un periodo definido por el programador, y después se transmiten los datos a un host centralizado u otros hosts que se encuentran distribuidos. Posteriormente, las otras bases de datos son actualizadas, de manera que todos los sistemas están en sincronía unos con otros. Este tipo de procesamiento distribuido ocurre generalmente entre computadoras departamentales o redes LAN y sistemas propietarios.

3.4.2.5 Lenguajes de Programación de Aplicaciones de Bases de Datos

Un DBMS sofisticado pierde importancia si los usuarios no tienen una manera de accesar a los datos.

Una aplicación de bases de datos es un programa que permite a los usuarios ingresar, modificar, borrar y obtener reportes de los datos que se encuentran en la base de datos. Las aplicaciones generalmente son escritas por los programadores, en uno o más lenguajes de programación especializados. Sin embargo, recientemente ha existido una tendencia al uso de herramientas de acceso a bases de datos orientadas al usuario que simplifican el proceso de uso de un DBMS y eliminan la necesidad de la programación a la medida.

Los lenguajes utilizados para crear aplicaciones de bases de datos pueden ser agrupados en tres categorías: lenguajes procedurales, lenguajes SQL y todos los demás lenguajes.

3.4.2.5.1 Lenguajes procedurales

La gran mayoría de los lenguajes de programación pueden ser descritos como procedurales. Cuando un programador crea una aplicación de bases de datos en uno de estos lenguajes, ha escrito el código de la aplicación como una serie de procedimientos. Cada procedimiento hace el trabajo de una porción de la aplicación, tal como una consulta a la base de datos o un procedimiento para modificar los datos contenidos en la base de datos. Los diferentes procedimientos son ligados unos con otros mediante otros procedimientos que forman la interfaz con el usuario (por ejemplo, un menú) y son ejecutados en un punto determinado de la aplicación.

Los lenguajes de programación estándar, tales como Pascal, COBOL, BASIC y C, son lenguajes procedurales. Estos lenguajes pueden ser utilizados para crear aplicaciones de bases de datos mediante el uso de una Interfaz de Programas de Aplicación (Application Programming Interface -API-), la cual consiste de un conjunto de funciones estándar (o llamadas) que apoyan al lenguaje para darle acceso a los datos de la base de datos. Las funciones del API generalmente se encuentran contenidas en "bibliotecas" que se encuentran incluidas en la aplicación cuando ésta es compilada. Varios vendedores de DBMS´s tienen estas bibliotecas disponibles como parte del paquete del DBMS o bien, como una opción adicional que implica un costo. Todos estos lenguajes de alto nivel, que además pueden ser utilizados para crear aplicaciones sin bases de datos, son referidas generalmente como "lenguajes de tercera generación (3GLs)".

Algunos lenguajes de programación procedurales son específicos para un DBMS particular. Estos lenguajes son referidos comúnmente como "lenguajes de cuarta generación (4GLs)". Los ejemplos más comunes de lenguajes procedurales para bases de datos específicas son: dBASE, PAL (Paradox Application Language) y el R/BASIC Language (utilizado por Advanced Revelation). [8]

3.4.2.5.2 Lenguajes de Consulta Estructurados (SQL -Structured Query Language -)

El Lenguaje de Consulta Estructurado (SQL) fue diseñado como un lenguaje para accesar explícitamente a DBMS´s basados en el modelo relacional. La versión inicial del lenguaje apareció como SEQUEL a mediados de los años 70´s y fue desarrollado por IBM como un lenguaje estándar que corría sobre mainframes IBM. Posteriormente, su nombre fue reducido a SQL.

SQL se describe más propiamente como un sublenguaje, ya que no contiene facilidades de manejo de pantalla o de entrada/salida. Su objetivo principal es proveer un método estándar para accesar a las bases de datos, independientemente del lenguaje en que se encuentre escrita el resto de la aplicación. Está diseñado para consultas interactivas de una base de datos o como parte de una aplicación escrita en algún lenguaje procedural. [9]

Desde que fue creado, ha sido revisado en distintas ocasiones; a principios de los 80´s la organización ANSI (American National Standards Institute) intentó estandarizar el lenguaje SQL, lo que condujo a la liberación de las especificaciones ANSI-86 SQL y posteriormente las especificaciones ANSI-89 SQL y ANSI-92 SQL. IBM se ha enfocado a la expansión del lenguaje SQL, trabajando especialmente en ampliar las capacidades de su base de datos relacional para mainframes DB2. Cada vendedor de DBMS incluye sus propias extensiones al estándar SQL, y estas extensiones pueden hacer a las distintas versiones de SQL incompatibles una con otra.

SQL fue inicialmente introducido comercialmente como un sistema de bases de datos en 1979 por Oracle.

En 1991, ANSI actualizó el estándar. El nuevo estándar es conocido como SAG SQL. [10]

Se han derivado una serie de lenguajes de consulta, todos basados en el SQL estándar, alguno de ellos son:

Lenguajes SQL
SQLDB: (Structured Query Language Database). Es un lenguaje creado para definir y manipular bases de datos relacionales como lo especifica la ANSI en el estándar X3.135-986 y el estándar X3.115-1989 de la FIPS (Federal Information Processing Standard).
Transact-SQL: Lenguaje nativo de Sybase, es una extensión del SQL - ANSI
SQL92: Lenguaje nativo de ORACLE, es el estándar más reciente de ISO **PL/SQL:** Es una extensión de SQL. **SQL*Plus:** Es un interprete de comandos SQL para diversos sistemas operativos.
ESQL/C: Creado por Informix. Hace una combinación de sentencias SQL y sentencias del lenguaje C.
MiniSQL: Desarrollado por David Hughes de Hughes Technologies. **MSQL:** Es un lenguaje de consultas para uso ideal con páginas html generadas a partir de consultas de bases.
Inquery: Es un lenguaje de consultas para bases de datos. Soporta dos tipos de consultas: Lenguaje natural y estructurado. Evalúa información de acuerdo a modelos probabilísticos, y no a modelos booleanos. Es decir encuentra documentos que concuerden con todas las palabras en la consulta y decide que documentos son los más adecuados. Inquery's es otro método, de consultas estructuradas, que permite obtener información más exacta acerca de las relaciones de términos en la consulta.
Query-By-Example (QBE): No es estrictamente un lenguaje; es una interfaz que presenta al usuario con una o más tablas en blanco que corresponden a las tablas en la base de datos. El usuario elige las columnas que serán incluidas en la consulta a través de una combinación de teclas, y define las condiciones de la consulta llenando las condiciones dentro de las columnas apropiadas. El DBMS traslada el QBE en las acciones necesarias para cumplir con el requerimiento del usuario. [11]

3.4.3 Datawarehousing y soporte de decisiones (1990s)

Un Datawarehouse o almacén de datos o bodega es usualmente definido como una colección de datos integral variable en el tiempo, no volátil y orientada a temas importantes para el soporte a la toma de decisiones en la administración de la organización.

Características:

1. Debe ser integral; no importando de donde provienen las fuentes de datos que lo integran deberán formalizarse como un solo tipo descrito de datos.
2. Debe estar organizado en relación al objetivo principal de la aplicación y las variables más importantes de la organización. Esta información es el soporte real para la toma de decisiones por lo que su objetivo particular debe ser servir a la misión central de la organización.

3. Debe ser incrementable en el tiempo. Debe almacenar información de varios años
4. No debe ser volátil. No debe permitir agregaciones modificaciones o eliminaciones de los datos que contiene cuando estas provengan del usuario.

Data warehouse es un concepto relativamente nuevo y desconocido en México y otros países, que viene a resolver los problemas de manejo y uso adecuado de grandes fuentes de datos y de muy diversos tipos, para apoyar la toma de decisiones oportunas y fundamentadas.

Un sistema operacional o de procesamiento en línea (OLTP: On-line Transaction Processing), es un sistema tal como el de administración de recursos humanos, de asignación de créditos bancarios, de recuperación y control de cartera o de control de seguros, y su función principal es dar el soporte a las necesidades diarias de la empresa; son sistemas normalmente optimizados para el manejo de un conjunto predefinido de transacciones.

Los sistemas operacionales de los cuales se transferirá la información seleccionada, pueden haber sido construidos utilizando manejadores de datos relacionales (RDBMS, del inglés), manejadores de archivos jerárquicos, de archivos planos u otro tipo de manejadores. Por lo anterior es necesario analizar y definir cuidadosamente de los sistemas operacionales aquellos datos que representen la esencia o filosofía del negocio que se pretenda manejar, para que al transferir los datos a la bodega, ese conocimiento primordial se capture en lo que se conoce como metadatos, que son precisamente los que describen a los datos provenientes de los sistemas operacionales. Los metadatos en general definen los formatos, significado y origen de los datos y facilitan, por lo tanto, el acceso, la navegación y la administración de los datos en la bodega.

Los datos extraídos de los diferentes sistemas pueden, a su vez, ser manejados por un RDBMS.

Data warehouse no es ni un producto de software ni una máquina o tecnología de bases de datos en particular, sino una serie de componentes y procesos que en conjunto forman la arquitectura data warehouse.

Sus partes más importantes son:

- ***Análisis, selección y extracción de datos.*** Procesos requeridos para seleccionar datos de sistemas operacionales, extraerlos y convertirlos a un formato o formatos que permitan manejarlos en común, de acuerdo al modelo de datos de la empresa, y de acuerdo a la información para toma de decisiones que se desee contar. Los datos deberán ser actualizados (extraídos de nuevo) como un proceso cíclico y periódico.

- ***Almacenamiento de los datos extraídos.*** Elementos necesarios para almacenar y manejar los datos. A este nivel se cuenta con los detalles de los datos, así como con los metadatos o información de alto nivel que los describe, obtenidos en el análisis del paso anterior. Para el almacenamiento y manejo de los datos extraídos se puede utilizar un RDBMS, el cual, por su manejo de grandes volúmenes de información, del

orden de terabytes incluso, por su implementación destinada a explotar las facilidades de hardware tales como memorias de 4 gigabytes o mayores y procesamiento simétrico o SMP, los hace eficientes para consultas, selección y procesamiento de datos.

- ***Análisis de los datos.*** Análisis simple. Explotación de los datos realizando consultas simples basándose en herramientas de productividad de oficina tales como hojas de cálculo o paquetes de análisis estadísticos.

- ***Análisis complejo.*** Explotación de los datos basándose en consultas y análisis multidimensional, utilizando herramientas de software para procesamiento analítico u OLAP (On-Line Analytical Processing).

Actualmente es innegable que los sistemas de información OLTP, construidos utilizando manejadores de bases de datos relacionales son la norma. Es una tecnología madura que provee las facilidades necesarias. Cuando los usuarios de negocio empujados por las necesidades del mercado iniciaron con sus demandas de información actualizada, de proyecciones en el tiempo, de análisis comparativos entre regiones en diferentes períodos, los desarrolladores utilizaban las herramientas y manejadores de bases de datos que tenían al alcance, esto es, RDBMS. Lo anterior explica que algunos autores asocien un manejador de bases de datos relacionales cuando se trata de manejar el data warehouse.

Los objetivos de los sistemas OLTP y data warehouse para toma de decisiones son muy diferentes; tratar de diseñar un data warehouse pensando en un sistema operacional es garantía de fracaso. El éxito en la implantación de una arquitectura de data warehouse en las empresas radica en parte en el éxito en el diseño de los sistemas operacionales, ya que éstos son los proveedores de datos y los que se deben adaptar rápida y flexiblemente a los cambios del negocio. Se deben tener ideas muy claras de lo siguiente: Qué datos se deben utilizar; cómo se deben transformar; cómo se deben transferir, almacenar y organizar; y finalmente, cómo se deben acceder y analizar.

Datawarehousing: es el proceso de extracción y transformación de datos obtenidos en fuentes operacionales (OLTP) llevándolos a una base de datos centralizada reconocida como un Datawarehouse. Por lo que una vez en el repositorio es explotada usando herramientas para la toma de decisiones. [5]

3.4.4 Minería de datos (Hoy en día)

A lo largo de la vida productiva de una empresa se van acumulando grandes cantidades de datos que son almacenados en algún lugar para posteriormente hacer uso de ellos.

Gracias a los sistemas de cómputo, las empresas tienen la capacidad de almacenar y accesar, en archivos o bases de datos, grandes cantidades de datos históricos sobre las operaciones diarias de su negocio; información que en su momento fueron usados para satisfacer las necesidades propias de la empresa y para el soporte de decisiones.

Todos esos archivos deben contener gran cantidad de información que sería de gran utilidad si fuera posible aprovecharla.

La mayoría de las organizaciones no sufre por falta de información, sino más bien por exceso de información redundante a la que resulta cada vez más complicado acceder para buscar datos específicos y significativos que permitan obtener una visión más completa de la situación operacional de la empresa y así lograr una mejor toma de decisiones.

Las áreas de sistemas han venido trabajando para crear extractos de información de las bases de datos operacionales y almacenar estos datos en archivos, tratando de responder a las peticiones de los usuarios para obtener información que les ayude a tomar mejores decisiones. Las necesidades de información han hecho que se diseñen sistemas de información ejecutiva y de apoyo a la toma de decisiones -- tienen como objetivo primordial proveer de toda la información necesaria a los ejecutivos de alto nivel para apoyarlos en la toma de decisiones, además de que les permite tener acceso rápido y efectivo a la información compartida y crítica del negocio --; sin embargo, las demandas de las empresas, en relación a la información, van más allá de simples consultas, cruces de información o reportes consolidados; lo que ha hecho que se creen nuevas formas de análisis de la información. [12]

El objetivo del presente trabajo de investigación es introducir al lector al tema de la minería de datos presentando sus características, estructura y aplicaciones para despertar el interés de la gente involucrada en el proceso de toma de decisiones, que deseen descubrir datos útiles en su información histórica y obtener un valor agregado de ésta.

Existen algunos tipos de empresas que su principal negocio es manejar la información; por lo que siempre tratan de convertirla por todos los medios posibles en conocimiento.

La minería de datos esta lista para aplicaciones de las empresas actuales ya que esta soportada por tres tecnologías que actualmente están lo suficientemente maduras:

- Colección masiva de datos
- Poderosas computadoras de multiproceso
- Algoritmos de minería de datos

3.5 Clasificación

La minería de datos esta clasificada, a grandes rasgos, en tres categorías:

1. Análisis estadístico o de datos: se usan para detectar patrones no usuales de datos, estos patrones de datos se explican después mediante modelos estadísticos y matemáticos. Algunas técnicas de modelado son el análisis lineal y no lineal, análisis de regresión continua y logística, análisis de variación y multivariación y el análisis de series históricas. Las herramientas de análisis estadístico se utilizan en varias aplicaciones empresariales: incrementar la participación en el mercado y las utilidades, determinando las mejores oportunidades; mejorar la satisfacción del usuario mejorando la calidad en productos y servicios.

2. Descubrimiento de conocimiento: consiste en un conglomerado de componentes que identifican y extraen patrones y relaciones interesantes y útiles. Las principales entradas al sistema de descubrimiento de conocimiento son: los datos del Data Warehouse, la guía del analista empresarial y los conocimientos en la materia que almacena la base de conocimiento del sistema.

3. Otros tales como la visualización y sistemas de visualización geográfica.

3.6 Características generales

- La minería de datos auxilia a los usuarios empresariales en el procesamiento de vastas reservas de datos para descubrir relaciones insospechadas.
- La información obtenida a través de la minería de datos ayuda a los usuarios de sistemas de información a elegir cursos de acción y a definir estrategias competitivas.
- Los seres humanos tienen agudeza para percibir excepciones y anormalidades, pero no tienen la potencia y capacidad para inferir relaciones en grandes volúmenes de datos, por lo que la minería de datos puede examinar gran cantidad de datos encontrando patrones los cuales son difíciles de detectar.
- Trabaja con grandes cantidades de información histórica.
- La minería de datos usa modelos avanzados y reglas de inducción para encontrar tendencias y patrones en los datos.
- El proceso de búsqueda puede ser hecho por herramientas las cuales automáticamente buscan patrones por si mismas y despliegan los tópicos importantes.

3.7 Estructura

3.7.1 Algoritmos o programas que buscan (mineros).

La minería de datos hace uso de programas de búsqueda (llamados mineros) son los que se encargan de detectar desviaciones, tendencias y patrones ocultos en los datos históricos.

Los mineros son programas pensados y creados por el usuario en los que hace uso de algunas de las diferentes técnicas para el análisis y explotación de los datos, tales como: cluster, asociaciones, clasificación, visualización, redes neuronales, algoritmos genéticos, detección de desviaciones, entre otros.

La función de los programas mineros es la correlación de los criterios de selección y búsqueda con los datos históricos, si encuentra algo interesante lo presenta al usuario como un hallazgo.

Los programas mineros trabajan, principalmente, sobre bases de datos relacionales como procesos automáticos; buscan datos extraños,[1] patrones, tendencias o desviaciones; pueden ser ejecutados fuera de las horas pico, usando tiempos de máquina excedentes de noche o en horas de poco proceso, lo que los convierte en ayudantes importantes que utilizan el mismo criterio que el tomador de decisiones.

Una ventaja de los mineros es que no requieren hardware especial o dedicado. Trabajan en las redes de oficinas nacionales (o regionales), utilizando por las noches el servidor de la base de datos relacional, y las PCs o estaciones de trabajo ya existentes. Es decir trabajan sobre datos ya recolectados, en máquinas ya existentes, realizando labores útiles mientras los usuarios duermen.

3.7.2 Datos históricos (en dónde buscan).

Son datos estables y coherentes en un momento dado que se van acumulando a lo largo de la vida operativa de una empresa.

3.7.3 Criterios de búsqueda (qué se buscan).

Son las normas, tendencias y patrones bajo las cuales los programas mineros realizarán el proceso de selección y búsqueda en los datos históricos. La prioridad de búsqueda, los criterios de interés y las definiciones de situación extrañas son definidas por el usuario. Una vez establecidos los criterios de selección y búsqueda, los mineros analizan los datos históricos reportando los hallazgos inmediatamente en un archivo para su posterior visualización y decisión final.

[1] Datos extraños. Datos desconocidos o poco comunes en la información que se esta manejando.

3.7.4 Almacenamiento de hallazgos (Cofre de tesoros).

Los hallazgos son los datos resultantes de correlacionar los criterios de selección y búsqueda con los datos históricos. El ser humano desempeña un papel fundamental, ya que sólo él, el analista, puede decidir si este patrón, tendencia o criterio tiene importancia, pertinencia y utilidad para la empresa. [5]

3.8 *Usuarios de la minería de datos*

Los principales usuarios de la minería de datos son:

- Analistas empresariales.
- Peritos en estadística.
- Profesionales en tecnología de la información que apoyan a los usuarios empresariales.

Usuarios de los resultados de la minería de datos

- Grandes firmas manufactureras: Target Marketing, tendencias de construcción y tendencias económicas.
- Compañías dedicadas a la venta de productos: pronóstico de ventas, tendencias de mercado, niveles de inventario, correlación de ventas por producto, perfiles de vendedores.
- Compañías bancarias: análisis de crédito y riesgo, predicción de las acciones y valores en el mercado, análisis financiero, perfil del cliente, tendencias en el mercado, etc.
- Compañías aseguradoras: detección de fraudes, análisis de calamidad, administración del riesgo, mercado directo, administración de agencias, perfiles de vendedores, aseguradoras, etc.
- Manufacturación: desarrollo de productos, investigación de mercado, estrategias.

3.9 *Áreas de aplicación*

- Áreas de Planeación, evaluación, apoyo a la toma de decisiones y otras donde la necesidad de análisis semi-exhaustivo sea extensa, debido al gran número de datos.
- Investigaciones naturales: usan la minería de datos para explotar recursos no explorados, para examinar grandes cantidades de datos para actividades de minería.
- Pronóstico del tiempo: Usan datos históricos para entender los patrones del tiempo predecir las tendencias del tiempo.
- Actividades sísmicas: compilan datos para entender actividades sísmicas futuras basándose en actividades pasadas de problemas de áreas.
- Investigaciones geológicas ecológicas: usan datos para comprender la historia de la tierra y las tendencias globales que afectan el planeta.
- Investigaciones médicas: usan casos de estudios históricos que determinan atributos de sugerencias y aplicaciones de datos actuales de los pacientes.
- Investigaciones académicas: usan la minería de datos para el desarrollo de métodos estáticos, teoría de análisis de datos y métodos, etc. [13]

3.10 Proveedores

Existe una gran cantidad de proveedores de software minero en la actualidad. Se presenta un cuadro sinóptico en el Anexo 3, en el cual se muestran los datos más generales de los proveedores existentes en el mercado y las características de la herramienta de extracción de datos que ofrecen.

3.11 Plataformas

3.11.1 Clasificación

Una **plataforma** es un conjunto formado por: hardware, software de sistemas (sistema operativo), software de subsistemas (Manejador de base de datos), software de aplicación y software de utilerías. Existen diversos tipos de plataformas las cuales se presentan a continuación:

Mainframes o minis enlazadas

Funcionan como servidor gigante de archivos o servidor de bases de datos, manejan aplicaciones tanto por lotes (batch) como interactivas y la interfaz con el usuario es de tipo carácter, además de que son equipos propietarios.

Minis o workstation de alto desempeño

Dan servicio a nivel departamental, pero a diferencia de las workstations, las minicomputadoras son equipos propietarios y también manejan aplicaciones tanto batch como interactivas. Son de muy alto costo y generalmente son fabricadas por empresas como Unisys, Hewlett Packard y Digital donde los sistemas abiertos no son su principal producto.

Workstation o PCs de alto desempeño

Fundamentalmente manejan aplicaciones interactivas (en línea), pueden funcionar aisladas (stand-alone) o interconectadas en red.

Microcomputadoras

Son computadoras personales que en relación a las anteriores tienen menor capacidad y poderío pero que gracias al avance de la tecnología cada vez se va acrecentando sus capacidades de almacenamiento y procesamiento. Son muy conocidas y ampliamente disponibles. Representa un alto costo la compra de PCs poderosas y bien equipadas para ser utilizadas como servidores de bases de datos dedicados.

3.11.2 Front-ends

Se denomina Front-end a la unidad central encargada de interconectar a un usuario con el sistema que atenderá su requerimiento. Es la computadora que recibe directamente las líneas de entrada. [6]

Los paquetes front-ends pueden ser divididos en cuatro grandes categorías basadas en su función principal: (add-on) para productos existentes, herramientas de desarrollo de aplicaciones, programas reporteadores/consultas y herramientas de análisis e integración de datos.

En la siguiente tabla se muestran algunos de los productos front-end existentes en el mercado [14].

Front End	Proveedor	URL
SELECT Component Factory™ (SCF)	SELECT Software Tools	http://www.selectst.com/Welcome.asp
Delphi	Borland International Inc.	http://netserv.borland.com/delphi/papers/delcsoa/#overview
Centura Team Developer	Centura Software Corporation (previously Gupta Corporation)	http://www.centurasoft.com/products/development/
The Forté Application EnvironmentTM	Forté Software	http://www.forte.com/product/utc/
AppMaster Builder Fact Sheet	INTERSOLV	http://www.intersolv.com/appmaster/builder/frameset_builder.html
Prolifics	JYACC	http://www.prolifics.com/prod/index.html
Client Development System	Magic Software	http://www.magic-sw.com/products.html
Visual Basic 5.0 Enterprise	Microsoft	http://exodus.dcaa.unam.mx/publica/dba/visual.html
The WCL/Client	Multi Soft, Inc.	**¡Error! Marcador no definido.**
Developer/2000 2.0	ORACLE	http://exodus.dcaa.unam.mx/publica/dba/visual.html
PowerBuilder	Powersoft	http://www.powersoft.com/products/powerbuilder/

3.11.3 Back - ends

Se denomina Back-end al procesador que se utiliza para una función especializada, tal como la administración de una base de datos o una unidad especializada de aritmética y lógica. Es la parte destinada a recibir las peticiones y solicitudes del cliente, se encarga de manejar el procesamiento de los datos y el acceso directo a disco.[8]

En el anexo 2 se presenta una tabla que contiene información general de los back-ends existentes actualmente en el mercado (Datos actualizados al 26 de Junio de 1998) [15].

3.12 Conclusiones

El antecedente de la minería de datos son las bases de datos.

La minería de datos es el resultado de un largo proceso evolutivo que va desde la captación y acumulación de datos a través del tiempo, el almacenamiento y acceso en diferentes medios y formas, hasta llegar al análisis y explotación de los datos para convertirlos en información útil que ayude al soporte de decisiones.

Los archivos, las bases de datos y el data warehouse han tenido el mismo objetivo que es: permitir al ser humano obtener información que le apoye en la difícil tarea de tomar decisiones certeras que le guíen al logro de sus objetivos.

En la evolución de la información nos encontramos a nuestro paso la creación de bancos de datos, bases de datos, data warehouse hasta lo que hoy en día se conoce como minería de datos.

Todos los avances en la información tienen como objetivo primordial proporcionar información que apoye al ser humano en la difícil tarea de tomar decisiones certeras que lo orienten hacia el logro de sus objetivos.

Minería de datos es el proceso cuyo objetivo es descubrir, extraer y almacenar información relevante y previamente ignorada de amplias bases de datos a través de programas de búsqueda de criterios definidos por el analista, en grandes cantidades de datos.

La minería de datos puede ser aplicada a la información de cualquier empresa independientemente de la forma de almacenamiento en la que esta se encuentre.

Con las bases de datos los programas mineros pueden llegar a tener un buen desempeño pero lo ideal es el datawarehouse para la extracción de la información histórica.

La minería de datos se ha convertido en una fuente importante de creatividad, sobre todo para aquellos que ofrecen herramientas auxiliares para los tomadores de decisiones, lo cual lo podemos observar en la gran cantidad de software minero existente en el mercado que se va acrecentando cada día más.

Marco metodológico

4. Marco Metodológico

4.1. Conocer la información de la empresa

El estudio detallado de la información histórica de la empresa es trascendental, ya que esto permite conocer la disponibilidad, confiabilidad, consistencia, flexibilidad y eficiencia de los datos y de la implementación física en la que se encuentran almacenados.

4.1.1. Conocer la forma de almacenamiento.

En el panorama industrial de hoy, las empresas han resaltado la importancia que tiene la información para la toma adecuada de decisiones que la lleva a elevar el nivel de servicio de los clientes, reducir sus inventarios, incrementar sus ventas, simplificar y mejorar el nivel de control contable, financiero y gerencial del negocio y, en general, a obtener mayores ganancias, lo que es el fin principal de la mayoría de las empresas.

Antes de comenzar a determinar lo que se desea obtener como salida es necesario conocer, al detalle posible, lo que funcionará como entrada en este caso los datos acumulados con el transcurso de los años por la ejecución de las operaciones normales de una empresa y que relacionados se convierten en información. Lo fundamental de este paso radica en la concienciación que hace el analista de los insumos con los que cuenta. En la medida que una empresa capte más datos de sus operaciones tendrá la capacidad de correlacionarlos según sus necesidades y así obtener mayor oportunidad de conocer su empresa. Lo anterior no quiere decir que las empresas deberían capturar todos los datos generados por su operaciones cotidianas sino que tendrá que estudiar con lo que cuentan y definir los datos, adicionales a los que ya tienen, que necesitan para comenzar con su captación.

Se deberá conocer las diferentes formas de almacenamiento de los datos de tal manera que sea posible visualizar el contenido de la información de los mismos con el mínimo de perturbaciones y así proporcionar un entendimiento sobre cómo los datos, están relacionados.

A continuación se describen las formas de almacenamiento más comunes de las que se hacen uso para la organización de los datos.

Los datos suelen estar dispersos en una gran cantidad de plataformas, sistemas operativos y programas diferentes.

4.1.1.1. Banco de datos

Al hablar de bancos de datos se ha considerado que estos existen totalmente independientes entre sí. En la realidad, el manejo de combinaciones de tales archivos resulta difícil por que los datos se almacenan en diferentes formatos y en distintos

archivos, una característica de los archivos es que los datos no se pueden compartir entre los diferentes programas, a menudo no son recuperables ni están seguros.

Los registros referentes al sistema se mantienen permanentes, además de estos archivos el sistema cuenta con varios programas de aplicación que permiten manejar los archivos. Estos programas fueron escritos por programadores en respuesta a las necesidades de la organización, y cada vez que se necesita se agregan nuevos programas de aplicaciones al sistema. Como resultado se crean nuevos archivos permanentes que contienen toda la información acerca de la empresa, por lo cual tal vez sea preciso escribir nuevos programas de aplicación. Puesto que estos archivos y programas se han creado para un período largo y, probablemente, por distintos programadores; es de esperar que los archivos tengan formatos diferentes.

La característica principal de los bancos de datos es el aislamiento, con en este tipo de almacén los datos están repartidos en varios archivos, y éstos pueden tener diferentes formatos, es difícil escribir nuevos programas de aplicaciones así como aplicar criterios de selección y búsqueda para obtener los datos apropiados.

Los problemas a los que se enfrenta un minero al trabajar con este tipo de almacén son: redundancia e inconsistencia de los datos, dificultad para tener acceso a los datos, aislamiento de los datos, problemas de seguridad, problemas de integridad, entre otros.

Para que un minero pueda trabajar de forma óptima es necesario trabajar con datos consistentes, no redundantes, seguros y confiables e integros. Problemas que los bancos de datos tienen debido a su aislamiento.

4.1.1.2. Base de datos

El manejo de los datos nunca es sencillo, aunque parece un problema sencillo, existe una serie de dificultades a vencer.

Debido a la importancia que tiene la información en casi todas las organizaciones, la base de datos en un recurso valioso ya que permite manejar los datos en forma eficiente y controlar problemas como redundancia e integridad, el accesos a los datos, problemas de seguridad, problemas de integridad y aislamiento.

Los sistemas de base de datos se diseñan para manejar grandes cantidades de información. El manejo bases de datos incluye tanto la definición de las estructuras para el almacenamiento de la información como los mecanismos para el manejo de la información.

Con el modelo de datos relacional se representa la base de datos como un conjunto de tablas que tienen asignado un nombre único. Una columna de una tabla representa una relación entre un conjunto de valores. Puesto que una tabla es un conjunto de estas relaciones, existe una correspondencia entre el concepto tabla y el concepto matemático relación, del cual recibe su nombre el modelo de datos relacional.

A través del modelo relacional se pretende capturar, parcialmente, el significado de los datos para obtener un conocimiento del entorno del negocio.

En sus términos más simples una relación denota algún tipo de comparación o conexión entre un par de tablas.

Como sabemos un sistema de manejo de base de datos (DBMS) consiste en un conjunto de datos relacionados entre sí y un grupo de programas para tener accesos a esos datos. El conjunto de datos se conoce comúnmente como base de datos. Ésta contiene información acerca del negocio. El objetivo primordial de un DBMS es crear un ambiente en que sea posible guardar y recuperar información de la base de datos en forma conveniente y eficiente.

Los DBMS introducen numerosos conceptos relevantes desde el punto de vista de lectura y manejo de los datos. Proporcionan comandos capaces de procesar múltiples registros a la vez. Provee a los usuarios finales para interactuar directamente con la base de datos, especialmente aquellas consultas no planeadas. Hace una distinción clara entre la visión de los datos que tiene el programador (lógica) y la manera cómo éstos se encuentran en los dispositivos magnéticos (física).

Al combinar las bondades de un almacén relacional y los criterios de selección y búsqueda de los mineros se tendrá la certeza de los siguientes conceptos:

- Independencia de los datos: Proporcionar una frontera clara y tangible entre los aspectos lógicos y físicos de un manejador de base de datos.
- Comunicación: Crear un modelo empleando estructuras simples, de tal suerte que todo tipo de usuario (programadores y usuarios finales) tenga una comprensión común de los datos, y por lo tanto, puedan comunicarse entre sí sobre los aspectos de la base de datos.
- Procesamiento de conjuntos. Introducir conceptos para un lenguaje de alto nivel que permita a los usuarios expresar operaciones de grandes volúmenes de datos, en lugar de tener que considerar un registro a la vez

4.1.1.3. Data warehouse

Como su nombre lo indica, el data warehouse (bodega o almacén de datos) es una enorme colección de datos provenientes de sistemas operacionales y otras fuentes, después de aplicarles los procesos de análisis, selección y transferencia de datos seleccionados. El objetivo principal del data warehouse es el uso adecuado de esos datos para obtener información útil para el soporte a la toma de decisiones, lo que es difícil de lograr con sistemas operacionales.

Un sistema operacional o de procesamiento en línea OLTP (On-Line Analytical Processing), es un sistema que su función principal es dar el soporte a las necesidades del día de la empresa; son sistemas normalmente optimizados para el manejo de un conjunto predefinido de transacciones.

Los sistemas operacionales de los cuales se transferirá la información seleccionada, pueden haber sido construidos utilizando manejadores de datos relacionales (RDBMS), manejadores de archivos jerárquicos, de archivos planos u otro tipo de manejadores. Por lo anterior es necesario analizar y definir cuidadosamente de los sistemas operacionales aquellos datos que representen la esencia o filosofía del negocio que se pretenda manejar, para que al transferir los datos al almacén, ese conocimiento primordial se capture en lo que se conoce como metadatos, que son precisamente los que describen a los datos provenientes de los sistemas operacionales. Los metadatos en general definen los formatos, significado y origen de los datos y facilitan, por lo tanto, el acceso, la navegación y la administración de los datos en la bodega.

El análisis y procesamiento de datos en el almacén se puede apoyar y completar con varias técnicas de análisis simples como la explotación de los datos realizando consultas sencillas basándose en herramientas de productividad de oficina, tales como hojas de cálculo o paquetes de análisis estadísticos; y análisis complejos como la explotación de los datos basándose en consultas y análisis multidimensional, utilizando herramientas de software para procesamiento analítico u OLAP; entre las que destaca la minería de datos.

El data warehouse representa una herramienta muy útil para el proceso de minería de datos, sin embargo muy pocas empresas cuentan con una estructura de este tipo.

Debido a que las bases de datos relacionales han sido ampliamente usadas por muchas empresas además de que proporcionan a la minería de datos grandes facilidades para las búsquedas y descubrimiento, hemos decidido orientar la presente metodología a la aplicación del proceso de minado de datos para bases de datos relacionales.

4.1.2. Conocer la estructura de almacenamiento de la información.

El analista antes de comenzar a definir sus criterios de búsqueda debe saber cuales son las partes que forman la estructura de su base de datos para así formular las posibles rutas de búsqueda necesarias para obtener los resultados planeados o bien ajustar sus criterios a la estructura de bases de datos con la que cuenta.

- Tablas

Las tablas se refieren a un arreglo de datos en renglones y columnas. En sistemas administradores de bases de datos toda la información esta almacenada en forma de tablas. El conocer la definición de las tablas permite conocer la estructura de la base de datos y ubicar la información para su posterior localización.

- Conocer las relaciones entre las tablas

Las relaciones entre las tablas están dadas por las llaves primarias y foráneas, las cuales denotan algún tipo de comparación o conexión entre un par de tablas. El conocer dichas relaciones permiten explotar mejor la información histórica, sobre todo conocer como están vinculados los datos.

- Identificar las dependencias entre las tablas

La dependencia es una propiedad de la información, que se representa con las relaciones. La dependencia no se determina con la diferencia de atributos en las relaciones o por el contenido actual de una relación. Tiene que ver con la dependencia de los valores de un atributo o un conjunto de atributos.

Es importante identificar las tablas "padre" para después ver cuáles y cuántas son las tablas "hijo"que dependen de ésta. Ayuda a conocer la estructura de la(s) base(s) de datos y las dependencias entre las tablas que la integran.

- Dominios

El domino para un atributo define los valores válidos que pueden tomar los atributos de cada registro. Así los dominios no son sólo tipos, como en los lenguajes de programación; son algo más abstracto. Dos series de valores pueden ser del mismo tipo pero de diferente dominio. Cada atributo se define como una serie de valores y cada atributo se asocia con un conjunto de estos dominios.

Un buen sistema relacional detecta un error cuando encuentra una expresión donde dos variables son del mismo tipo pero se derivan de diferentes dominios.

- Claves de relaciones o llaves primarias

El término llave primaria se define como una serie de atributos, cuyo valor identifica un renglón único en una relación. Así, dos renglones no pueden tener el mismo valor de clave, lo que a veces se llama propiedad de unicidad.

La importancia de las llaves primarias esta en identificar cada una de ellas dentro de las tablas involucradas en la base de datos y ver si alguna de estas se relaciona o no con alguna otra base de datos. Esto también es para seleccionar los campos clave para su futura explotación.

- Claves extranjeras o llaves foráneas

Una llave foránea es un atributo cuyo valor permite relacionar varias tablas. Su función es la conservación de la integridad de la base de datos.
Es indispensable identificar las relaciones existentes entre las tablas de las bases de datos. Su buen diseño ayuda a obtener mejores resultados en la definición de los criterios de selección y búsqueda.

- Tipo de datos

Es la clasificación de un tipo particular de información. Esto facilita a los humanos distinguir entre diferentes tipos de datos. El usar tipo de datos proporciona consistencia entre las columnas que están repetidas en diferentes tablas.

Conocer el tipo de información almacenada permite manejar apropiadamente la información y a determinar si se puede o no realizar operaciones en la información.

- Longitud

Conocer el tamaño de las columnas y del registro ayuda a hacer una estimación del tamaño total, en bytes, de cada una de las tablas y de la base de datos.

Esto es un buen punto de referencia para medir la cantidad de información histórica almacenada y poder hacer una mejor optimización del espacio ocupado en disco.

- Campos calculados

Son campos creados a partir de otros campos. El identificar dichos campos así como sus campos de procedencia facilita la comprensión del flujo de los datos de las bases de datos operativas.

- Restricciones (Constraint)

Son una alternativa para imponer integridad en las tablas de bases de datos. Integridad referencial significa que los valores de los atributo primos (llaves primarias) y los valores correspondientes de los atributos externos (llaves foráneas) deben coincidir de forma exacta. Permite saber si la información con la que se cuenta es confiable.

- Reglas

Una regla puede especificar una mascara de edición para una columna o un tipo de dato, puede ser una lista de valores, un rango o un patrón. Conocer si la información histórica tiene patrones de datos definidos.

- Defaults

Es el valor que se asigna por omisión a un campo de una tabla en específico cuando éste no es ingresado por el usuario. Una columna sólo puede tener un valor por omisión. Es importante conocer sí la información histórica tiene valores asignados por omisión.

- Índices

Los índices permiten obtener un acceso más rápido y eficiente de los datos. Es importante identificar como se hace el ordenamiento de los datos físicamente, si es lógico, con apuntadores, o físico en disco; facilitaría mucho la búsqueda de la información.

- Conocer el dato.

Implica saber exactamente qué información se almacena en cada campo. Es importante comprender el concepto asociado a ese dato y determinar su valor, concretar si es un dato básico o un dato complementario para su posterior uso. Es necesario conocer también el conjunto de valores válidos que puede tomar cada campo y la composición del registro.

Se hace una distinción entre los términos dato e información. Se emplea el término dato para referirse a los valores registrados físicamente en la base de datos, e información para aludir al significado de esos valores según el sentido que les dé un usuario.

4.1.3. Conocer y determinar el volumen de la información.

Los sistemas operacionales generan grandes cantidades de información proveniente de todos los departamentos que componen la empresa, misma que puede estar almacenada en diferentes medios, formatos, lugares o puede también, ser procesada

en diferentes plataformas. De hecho este es uno de los principales problemas a los que nos enfrentamos para la integración de la información.

A lo largo de la vida productiva de la empresa ésta información se va acumulando en grandes volúmenes y se va convirtiendo en información histórica. El volumen de información de una empresa está en función del número de transacciones registradas diariamente y por el tiempo que lleva laborando.

Antes de comenzar a trabajar con la información histórica es necesario seleccionar datos de los sistemas operacionales, extraerlos y convertirlos a un formato o formatos que permitan manejarlos en común, de acuerdo al modelo de datos de la empresa, y de acuerdo a la información para la toma de decisiones que se desee contar.

Es posible seleccionar información que corresponda a periodos muy grandes y de todas las bases de datos o se puede seleccionar sólo aquella que es generada por un sistema departamental. En general, conforme mayor sea el conjunto de información considerada mayores serán los resultados obtenidos de su análisis, pero también crecerá su complejidad en el mismo.

Los datos pudieron haber sido actualizados como un proceso cíclico o periódico; por ejemplo, cada semana, cada dos semanas o cada mes, de acuerdo a las necesidades de información actualizada que el negocio requiera. Tener o pretender análisis diarios es contraproducente en la mayoría de los casos, ya que la idea es trabajar con datos históricos con los cuales se puedan realizar comparaciones, proyecciones, etc. Tratar de comparar datos de un día con el anterior no es una pregunta para toma de decisiones y en caso de serlo, no se necesita un minero para obtener la respuesta ya que el sistema operacional sería suficiente.

4.2. Preparar la información para el proceso de minería de datos

Una vez que se tiene consciencia de la estructura y del volumen de la información histórica se procede a preparar la información para la toma de decisiones iniciando con la identificación de los problemas del negocio y las áreas en donde los datos pueden dar valor agregado a la empresa.

A raíz de un problema surge la necesidad de analizar a detalle los datos de la empresa para encontrar posibles soluciones al mismo, o bien, información que haga que las decisiones que sean tomadas sean lo más certeras posibles. Así mismo, es importante identificar las áreas donde la información es muy cambiante y que es primordial para la competitividad de la empresa. Es conveniente seleccionar una muestra de información para predecir debido a que es más sencillo que trabajar con toda la información.

Los problemas mal definidos tienen pocas probabilidades de dar resultados satisfactorios.

Para esto pueden identificarse varios criterios, no se puede decir específicamente cuáles son los correctos debido a que esto depende de las características de la empresa, pero el objetivo a perseguir es determinar los criterios, normas, ideas o

cuestionamientos que fungirán como entrada para el proceso de minado de datos. Con el trabajo normal de un sistema operacional muchas veces estos cuestionamientos no son respondidos por lo que se tiene que dedicar un proceso específicamente para obtener información más estudiada y precisa.

Es importante transformar la información convirtiéndola de un tipo a otro (valores nominales a valores numéricos), o bien definiendo nuevos atributos derivados (mediante la aplicación de operadores lógicos), a fin de que la información se adapte al análisis a efectuar. Dicha transformación no debe distorsionar el contenido de la información seleccionada. Puede ser necesario acceder a información adicional y realizar otras transformaciones de la información original. Pasará por una serie de posibilidades a medida que afine sus muestras de información para adaptarlas al carácter de los patrones ocultos en la información.

4.3. Usar los programas mineros para buscar en la información seleccionada los criterios, ideas, normas o cuestionamientos definidos.

En seguida se utiliza el algoritmo o algoritmos adecuados de minería para analizar la selección de datos. Estos algoritmos pueden hacer uso de modelos tales como: cluster, asociaciones, clasificación, visualización, redes neuronales, algoritmos genéticos, detección de desviaciones, entre otros.

Lo que todos estos algoritmos tienen en común es que requieren bases de datos de un tamaño considerable.

Los algoritmos seleccionados son traducidos a programas mineros que realizarán las búsquedas con los criterios definidos en el paso anterior.

Existen varias dificultades que pueden interferir con el resultado que se obtenga del análisis y esto es a razón de que los datos se pueden encontrar en diferentes formas, formatos y en múltiples sistemas, aunado a que, pueden provenir de fuentes internas o externas. Es muy importante tener claro que la información a "minar" deberá estar lo más uniforme y congruente posible ya que mucho depende de esto la certidumbre de los resultados que arroje.

4.4. Incorporar la información obtenida al proceso de toma de decisiones.

La información extraída, el cofre de tesoros, se analiza en función de los problemas identificados y de los objetivos de la empresa. Los resultados obtenidos son aplicados a la solución de los problemas reales definidos; la finalidad de la interpretación de resultados no es únicamente visualizar el resultado de la extracción; si no también filtrar la información que va a presentarse.

Presentar los hallazgos encontrados a los responsables de las operaciones de forma que los conocimientos obtenidos puedan integrarse en los procesos de la empresa y puedan ser aplicados en la solución de los problemas.

4.5. Medir los resultados

Una vez aplicados los resultados en forma de soluciones a los problemas definidos y al proceso de toma de decisiones se puede medir el valor agregado que da la minería de datos a la empresa.

Mientras se efectúa un paso determinado, a menudo es necesario revisar los pasos realizados previamente. Tras mostrar y medir los resultados de un hallazgo puede ser necesario preparar la información adicional, en cuyo caso es necesario regresar a preparar nuevamente la información para el proceso de minería de datos.

4.6. Ejemplos de aplicaciones de extracción de datos

Financiera
Al integrar información demográfica, de nivel de vida y de las cuentas de sus clientes, un banco podría utilizarla para descubrir clientes potenciales para sus fondos de inversión. Las bases de datos podrían dividirse con arreglos a criterios como residencia urbana, estado civil, ingresos, edad, bajo riesgo y alta liquidez. Luego podría lanzarse una campaña promocional para comercializar el nuevo producto entre cada grupo, adaptando el mensaje a sus distintas necesidades.

Mercadotecnia (Análisis de clientes fieles)
La extracción de datos puede utilizarse para el análisis de vulnerabilidad de clientes. El análisis de vulnerabilidad predice la fidelidad respecto a un producto o una clase de productos. Utilizando las predicciones de un modelo, las empresas pueden determinar a qué consumidores debe ir dirigida una estrategia de mercadotecnia determinada.

Aseguradoras (Análisis de siniestros)
La necesidad de reducir costes y la creciente competencia han hecho que las aseguradoras revisen reclamaciones que reciben para obtener una mayor eficiencia organizativa y para medir el grado de satisfacción de sus clientes. La comprensión de los factores que determinan el periodo necesario para liquidar un siniestro puede reducir el tiempo medio para tramitarlo.

4.7 Conclusiones

Los programas mineros que utilizan información almacenada en archivos uniformes o bases de datos operacionales no requieren de un almacén de datos. Sin embargo, la información situada en un almacén resulta más útil, debido a que la información se "limpia" antes de su almacenamiento.

Hasta hace poco, para extraer información de los almacenes se utilizaban únicamente generadores de consultas y sistemas de interpretación de información. En dichos casos se formula una hipótesis que se convierten una consulta y se coteja con la información seleccionada. La información así obtenida se analiza e interpreta.

Los sistemas que efectúan esta operación, denominados sistemas de verificación, presentan dos problemas. Primero, la persona que debe tomar la decisión debe conjeturar qué información necesita. Segundo, el valor de los resultados está limitado por la calidad de la interpretación. Si bien este análisis es adecuado para casos sencillos, la extracción por verificación no lo es para asistir a la toma de decisiones debido a la complejidad de la información operativa de las empresas y sus interrelaciones. Lo que se requiere es poder descubrir información importante oculta, para lo cual se utiliza el proceso de minería de datos.

La minería de datos combina funciones de verificación y de descubrimiento. La prospección mediante verificación permitirá expresar y verificar información en el ámbito organizativo y de personal. La prospección mediante descubrimiento se utilizará para refinar esta información y encontrar información sobre la que no se había avanzado una hipótesis.

La minería de datos es más que simplemente crear una base de datos y consultarla. A medida que la empresa obtiene mayor información sobre sus clientes, dicha información le permite adaptar su forma de hacer sus negocios a los deseos de sus clientes. Encontrar unos cuantos hechos clave en grandes montones de información - y concebir las ideas necesarias para comprender su importancia - puede resultar una tarea formidable que requiere de un profundo análisis, pensamiento claro y paciencia.

La metodología de minería de datos propuesta pretende orientar al analista empresarial en el proceso de minado de datos, sin embargo la definición de los criterios, patrones y tendencias a buscar en los datos históricos, así como las herramientas de análisis seleccionadas dependerán de la consistencia e integridad de los datos, de la visión que se quiera tener de ellos y de la arquitectura de sistemas de la empresa.

Marco instrumental

5. MARCO INSTRUMENTAL

5.1 Propuestas de acción

Es importante considerar que en este trabajo de investigación se sume al conocimiento, hasta hoy existente, del tema minería de datos. Conocimiento que contribuirá a que los empresarios de hoy en día descubran el valor que tiene su información histórica y que les puede encaminar hacia una mejor toma de decisiones.

Para tal efecto se decidió emprender los siguientes caminos acción.

- Publicación de un artículo en revistas.
- Diseño de una hoja html de minería de datos.
- Enviar el marco metodológico como folleto a universidades.

A continuación se desglosa a detalle cada una de estas actividades.

5.1.1 Publicación de un artículo

Con la publicación de un primer artículo se pretende dar una introducción de la investigación desarrollada, nuestro objetivo es despertar el interés de los lectores y de las personas administradoras de la información para presentarles una alternativa en búsquedas complejas e imprevistas; así como proporcionar una visión más amplia de lo que se puede hacer con la información y cómo ésta puede ayudar en la toma de decisiones.

En esta fase se seleccionaron 4 revistas de las cuales dos son especializadas en informática, una en administración, y la última en investigación en el área contable y administrativa. Las revistas seleccionadas son las siguientes:

Nombre	Giro	Editorial
PC Computing en Español	Computación e Informática	Editorial Ziff-Davis, S. de R. L. de C. V.
Byte en Español	Computación e Informática	Editora y comercializadora de bienes de informática S.A. de C.V.
Emprendedores	Al servicio de la pequeña y mediana empresa	Publicaciones Sayrols, S.A. de C.V.
Contaduría y Administración	Investigaciones de la FCA-UNAM	FCA-UNAM

5.1.2 Diseño de una hoja html de minería de datos

El objetivo es difundir la información de la minería de datos en español, con ligas a información ya existente en otros idiomas. Teniendo la posibilidad de contactar con otras personas interesadas en el tema.

Servidor WEB	Dirección
Drbaz	132.248.55.65

5.1.3 Enviar el marco metodológico como folleto a universidades

El marco metodológico tiene como propósito el dar a conocer la esencia de nuestra aportación en la investigación en forma resumida. Después de haber indagado sobre el tema "minería de datos", en dicho marco hemos descrito una metodología para la implantación del proceso de minería de datos. Esta metodología es el resultado del conocimiento que hemos obtenido a lo largo del desarrollo del trabajo y hemos considerado importante divulgarla para que otras personas lo conozcan y hagan uso de él si les es de utilidad.

Universidad	e-mail
ITESM, Campus Ciudad de México	biblio@campus.ccm.itesm.mx
ITAM, Instituto Tecnológico Autónomo de México	**¡Error! Marcador no definido.**
UAM, Universidad Autónoma Metropolitana	**¡Error! Marcador no definido.**
UG, Universidad de Guadalajara	A través de su página en Internet
UDLA, Universidad de las Américas – Puebla	**¡Error! Marcador no definido.**
UACAM, Universidad Autónoma de Campeche	Webmaster@jaina.uacam.mx
IPN, Instituto Nacional Politécnico	Internet@ipn.mx
UP, Universidad Panamericana	A través de su página en Internet
UNITEC, Universidad Tecnológica de México	Web@unitec.mx
UAA, Universidad Autónoma de Aguascalientes	Wwwadm@correo.uaa.mx

5.2 Plan y programa de trabajo

Las propuestas de acción mencionadas anteriormente fueron llevadas a cabo y a continuación presentamos los resultados de cada una de ellas, así como el tiempo establecido para cumplir con su objetivo.

- Publicación de un artículo.

PC COMPUTING en Español

Persona consultada:	Mercedes Galindo Favela (Director editorial)
Telefóno(s):	261 26 16 y 2 61 26 12
Fax:	261 27 32
Fecha:	20/Mayo/1998
Resultado:	"La revista PC COMPUTING no publica este tipo de artículos ya que los artículos publicados son traducciones por sus colaboradores de la revista PC COMPUTING (edición en inglés)."
Comentario:	Esta por enviarse curriculas.

BYTE en Español

Persona consultada:	Laura Mayo (Jefe de edición)
Telefóno(s):	605 99 62
Fax:	605 00 56
Fecha:	18/Abril/1998
Resultado:	"La revista BYTE sólo publica artículos de gente con 5 ó más años de experiencia comprobable, para publicar artículos que no cubran este requisito es necesario someter a evaluación del Consejo Editorial el contenido del artículo."
Comentario:	Se envió sinopsis del contenido del artículo para ser sometido a evaluación por el Consejo Editorial de la revista.

Emprendedores

Persona consultada:	Lic. María del Carmen Márquez
Telefóno(s):	622 83 96
Fecha:	22/Abril/1998
Resultado:	"La revista EMPRENDEDORES sólo publica artículos únicos."
Comentario:	Se recogió el artículo enviado.

Contaduría y Administración

Persona consultada:	Lic. Jorge Ríos Szalay (Jefe de la División de Investigación).
Telefóno(s):	622 84 75 / 65
Fecha:	22/Abril/1998
Resultado:	"El artículo se sometió a una minuciosa evaluación por parte del comité de la División de Investigación."
Comentario:	El artículo se publicará en la revista No 190 que corresponde al mes de julio, agosto y septiembre del presente año.

- Diseño de la hoja html.

La información esta en la siguiente dirección:

Nombre del servidor:	drbaz.fmedic.unam.mx
Responsable:	Ing. Martín de Jesús Jiménez
Administrador del servidor WEB	Ma. Isabel Angeles Larrieta
URL:	http://132.248.55.65/dataminig/index.html
Resultado:	Se aceptó la solicitud enviada al responsable, solicitando el espacio y los permisos requeridos para la creación de las hojas html.
Comentario:	Esta información estará disponible a partir del día 1 de julio de 1998.

- Enviar el marco metodológico como folleto a universidades.

Tratamos de establecer comunicación con las universidades mencionadas de las cuales únicamente recibimos respuesta de el ITESM.

Persona contactada:	Sr. Angel Fernando Blanco Soto
Universidad:	ITESM, Campus Ciudad de México
e-mail	biblio@campus.ccm.itesm.mx
Fecha:	22/Abril/1998
Resultado:	"El folleto será enviado a más tardar el 10 de Julio del presente, así se acordó con el Sr. Angel Fernando Blanco Soto."

5.3 Conclusiones

Con la publicación del artículo "Minería de datos: concepto, características, estructura y aplicaciones; una introducción a la extracción de datos útiles a partir de los ya existentes, para beneficio de las empresas" se logró difundir una parte de lo que es la minería de datos a la comunidad universitaria.

La página elaborada en Internet tiene un alcance más amplio y su objetivo es divulgar la información recopilada a lo largo de nuestro trabajo de investigación.

Consideramos importante compartir el conocimiento adquirido en la investigación realizada de la minería de datos a Universidades, sólo una de ellas envió respuesta.

Con las acciones tomadas en la implantación del presente trabajo de investigación queda cubierta la misión de difundir nuestro trabajo para que otras personas interesadas, en la minería de datos, puedan apoyarse en futuras investigaciones.

6. Conclusiones generales

En el gigantesco y complejo mundo empresarial se entrelazan varios factores que combinados eficientemente pueden llevar a una empresa a lograr el objetivo que desea, algunos de estos factores que logramos identificar al realizar el presente trabajo de investigación son: la organización de las partes componentes de la empresa y la información. La organización, por que de ésta depende, en gran medida, la buena comunicación entre departamentos o áreas, y la información por que esa comunicación entre departamentos y áreas genera precisamente información; que de ser información consistente, real y coherente puede servir de mucho para los tomadores de decisiones.

A la fecha, son pocas las empresas que han integrado exitosamente un sistema de información global para generar consultas, reportes y facilitar el análisis para la toma de decisiones. Los Sistemas de Información Ejecutiva (SIE) fueron soluciones populares en los 80's, las que se caracterizaron por ser simples y poco flexibles, ya que el enlace entre datos y resultados reportados no permitían una exploración rápida y con el detalle necesario. Para problemas específicos, se han desarrollado aplicaciones a la medida, con herramientas de Sistemas de Soporte para las Decisiones (SSD) y Bases de Datos Dimensionales (MDD) o alternativamente, con herramientas de productividad personal (hojas electrónicas de cálculo), que debido a su capacidad limitada, alto grado de mantenimiento y sin posibilidad de compartirse, eventualmente fracasan en su objetivo. El término minería de datos (Data mining) se ha propuesto como una solución absoluta a los problemas mencionados.

En el pasado, los mercados eran menos dinámicos, la competencia menor y menos agresiva, los márgenes de ganancias eran altos y los recursos humanos amplios. Los negocios se administraban mediante un estrecho contacto entre clientes y proveedores, y los administradores confiaban en su intuición y sentido común para tomar decisiones. La intuición y la "prueba y el error", ya no son métodos efectivos para administrar una empresa. La competencia se ha vuelto agresiva, los márgenes de utilidad son menores y el "staff" administrativo se reduce. Dada esta volatilidad, los empresarios se encuentran más en alerta sobre oportunidades que proporcionen un valor adicional a sus inversiones.

Se ha acumulado una gran cantidad de datos a través de sistemas tradicionales que administran las operaciones diarias de la empresa. En esta acumulación de datos, existe información con el potencial de incrementar participaciones en el mercado, mejorar la productividad, incrementar el rendimiento de la inversión y mejorar el servicio al cliente.

El concepto de minería de datos ha surgido a partir de la necesidad de los administradores de la información de aprovechar mejor la información que han ido almacenando y es precisamente la minería la que les permite descubrir aquellos datos valiosos que les pueden auxiliar en el proceso de toma de decisiones, proceso que encamina a las empresas hacia su éxito o fracaso.

Una serie de factores ha contribuido a poner de relieve el proceso de minería de datos entre los círculos empresariales los que a continuación se presentan:

- La aceptación general de que en las grandes bases de datos existen valores desaprovechados.
- Una consolidación de entradas de la base de datos que tiene una única presentación para el usuario.
- Una consolidación de las bases de datos, incluyendo el concepto de una almacén de información.
- Una reducción del coste de almacenamiento y procesamiento de datos que permite recopilar y acumular datos detallados.
- La intensa competencia por obtener la atención de los clientes en un mercado cada vez más saturado.
- La extracción de los datos es un proceso que permite no sólo obtener más información, sino información más profunda.

Hoy en día existen en el mercado una gran variedad de software minero encaminado a la extracción de datos, a la generación de patrones, tendencias y predicciones. Los usuarios de éstas herramientas se han enfrentado a descubrimientos que pueden ser falsos, inválidos o triviales cuando buscan patrones difíciles de detectar, pero eso no los ha detenido para buscar piezas de información que pudieran darles grandes ventajas sobre sus rivales.

Sin lugar a dudas el elemento más importante en el proceso de minería de datos es el ser humano, ya que es el único que analiza y decide en qué medida los hallazgos encontrados por la minería de datos le serán de utilidad.

El objetivo del presente cuestionario es recabar información acerca de nuestro tema de tesis, agradecemos de antemano su colaboración.

1. ¿En la organización en la que labora, utilizan algún modelo que les ayude al diseño de la información?

Sí ( ) No ( )

2. Si contestó afirmativamente la pregunta anterior, señale con "X" que modelos usa.

( ) Modelo relacional
( ) Modelo de red
( ) Modelo jerárquico
( ) Modelo distribuido
Otros. _____

3. ¿En la organización en la que labora, diseñan las bases de datos para su explotación inmediata o futura?

Inmediata ( ) Futura ( ) Ambas ( )

4. ¿Usa modelos que le ayuden al análisis de su información histórica?

Sí ( ) No ( )

5. Si contestó afirmativamente la pregunta anterior, señale con "X" que modelos usa.

( ) Clasificación
( ) Asociación
( ) Árboles de decisiones
( ) Redes neuronales
( ) Algoritmos paralelos
( ) Híbrido (Basado en arquitectura de redes neuronales)
( ) Reglas de inducción
( ) Estadísticas
( ) Análisis fractal
( ) Clustering
( ) Sequencias
( ) Algoritmos genéticos
( ) Memoria basada en razonamiento
( ) Reglas de clasificación
( ) Patrones generales y búsqueda de excepciones
( ) Sistemas de visualización
( ) Técnicas difusas
( ) Sistemas de información geográfica
Otras: _____

6. ¿Los modelos que usa le permiten hacer análisis exhaustivos de su información histórica de manera eficiente?

Sí ( ) No ( )

7. ¿Usa herramientas automáticas orientadas a la generación de información que le ayuden en la toma de decisiones?

Sí ( ) No ( )

8 Si contestó afirmativamente la pregunta anterior, marque con una "X" qué herramientas usa.

( ) DataCruncher
( ) DataMind
( ) AgentBase
( ) Business Miner
( ) DataBase Mining Marksman
( ) DataEngine
( ) MineSet
( ) Relationship Manager
( ) DecisionMaster
( ) Decision Series
( ) Ultragem Data Mining
( ) SuperQuery
( ) ModelMAX
( ) Data Surveyor
( ) Alice, AC2
( ) Management Discovery Tool (MDT
( ) DataDetective
( ) Syllogic Datamining Tool/MP
( ) Decisionhouse
( ) Otras:

9. ¿Considera que la información generada por sus sistemas actuales le permite conocer a detalle su negocio?

Sí ( ) No ( )

10. ¿Considera que el utilizar herramientas que le permitan generar información adicional a la que obtiene de sus sistemas, le puede ayudar a incrementar la productividad de su empresa?

Sí ( ) No ( )

11. ¿Conoce el término "minería de datos"?

Sí ( ) No ( )

Nombre: _____
Puesto: _____
Nivel académico: _____
Fecha: _____

Anexo 2
Back Ends

Back-End	Proveedor	Dirección	URL
Adabas	Software AG Americas Inc.	Reston, VA; 888-724-2394	www.sagafyi.com
Adaptive Server Anywhere	Sybase Inc.	Emeryville, CA;	www.sybase.com
Adaptive Server.	IQ Sybase Inc	Emeryville, CA; 800-879-2273	www.sybase.com
Angara Main Memory Database	Angara Database Systems Inc.	Palo Alto, CA; 714-622-5446, 650-322-1810	www.angara.com
Cach	InterSystems Corp.	Cambridge, MA; 617-621-0600	www.intersys.com
CodeBase	Sequiter Software Inc.	Edmonton, Alberta Canada; 403-437-2410	www.sequiter.com
CQL++ - SQL Data Management System	Tache Group Inc.	Melbourne, FL; 888-252-6050, 407-768-6050	www.tachegroup.com/products/CQL/takecontrol.html
D3	Pick Systems	Irvine, CA; 800-367-7425	www.picksys.com
DB2 Database Server 4	IBM Corp.	White Plains, NY; 800-426-4968, 520-574-4600	www.ibm.com
DB2 for MVS/ESA 4	IBM Corp.	White Plains, NY; 800-426-4968, 520-574-4600	www.ibm.com
DB2 Parallel Edition	IBM Corp.	White Plains, NY; 800-426-4968, 520-574-4600	www.ibm.com
DB2 Universal Database	IBM Corp.	White Plains, NY; 800-992-4777	www.ibm.com
Empress DB Server	Empress Software Inc.	Greenbelt, MD; 301-220-1919	www.empress.com
Empress RDBMS	Empress Software Inc.	Greenbelt, MD; 301-220-1919	www.empress.com
Extended Parallel Option	Informix Software Inc.	Menlo Park, CA;	www.informix.com
FairCom Database Server	FairCom Corp.	Columbia, MO; 573-445-6833	www.faircom.com
Heuristic Optimized Processing System	HOPS International Inc.	Miami Lakes, FL;	www.hops.com
IBM RS/6000 SP	IBM Corp.	White Plains, NY; 800-426-4329, 415-855-4329	www.ibm.com www.informix.com
Informix Dynamic Server	Informix Software Inc.	Menlo Park, CA;	
Informix Dynamic Server Developer Edition	Informix Software Inc.	Menlo Park, CA;	www informix.com
Informix Dynamic Server Workgroup Edition	Informix Software Inc.	Menlo Park, CA;	www.informix.com
Informix-DataBlade Modules	Informix Software Inc.	Menlo Park, CA;	www.informix.com
Informix-SE	Informix Software Inc.	Menlo Park, CA; 800-331-1763, 415-926-6300	www.informix.com

Back Ends

Back-End	Proveedor	Dirección	URL
InterBase	Borland International	Scotts Valley, CA; 888-345-2015, 408-430-1500	www.borland.com
Microsoft SQL Server Enterprise Edition	Microsoft Corp.	Redmond, WA; 800-426-9400, 425-882-8080	www.microsoft.com
Model 204	Computer Corporation of America	Framingham, MA;	cca-int.com
MSM for Windows NT	Micronetics Design Corp.	Rockville, MD; 800-258-2605, 301-258-2605	www.micronetics.com
MSM-PC/PLUS	Micronetics Design Corp.	Rockville, MD; 800-258-2605, 301-258-2605	www.micronetics.com
MSM-SQL	Micronetics Design Corp.	Rockville, MD, 800-258-2605, 301-258-2605	www.micronetics.com
MSM-Unix	Micronetics Design Corp.	Rockville, MD; 800-258-2605, 301-258-2605	www.micronetics.com
NonStop SQL/MP	Tandem, a Compaq Company	Cupertino, CA; 408-285-6000	www.tandem.com
OpenIngres	Computer Associates	Islandia, NY;	www.cai.com
Oracle Rdb	Oracle Corp.	Redwood Shores, CA; 800-542-1170, 650-506-7000	www.oracle.com
Oracle8 Database Server	Oracle Corp.	Redwood Shores, CA;	www.oracle.com
Pervasive.SQL	Pervasive Software	Austin, TX; 800-287-4383	www.pervasive-sw.com
Raima Database Manager++	Raima Corp.	Seattle, WA; 800-327-2462, 206-557-0200	www.raima.com
Red Brick Warehouse	Red Brick Systems Inc.	Los Gatos, CA; 800-621-2808, 408-399-3200	www.redbrick.com
Spatial Database Engine	Environmental Systems Research Institute Inc.	Redlands, CA; 800-447-9778, 909-793-2853 x11640	www.esri.com
Spatial Query Server (SQS)	Vision International of Autometric Inc.	Springfield, VA; 703-923-4300	www.autometric.com
SpatialWare	MapInfo Corp.	Troy, NY; 800-627-8627	www.mapinfo.com
SQLBase	Centura Software Corp.	Redwood Shores, CA;	www.centurasoft.com
Sybase Adaptive Server	Sybase Inc.	Emeryville, CA; 800-879-2273, 510-596-3500	www.sybase.com
System 1032	Computer Corporation of America	Framingham, MA;	cca-int.com
Teradata Parallel Relational Database	NCR Corp.	Dayton, OH;	www.ncr.com
Titanium	Micro Data Base Systems Inc.	West Lafayette, IN; 800-445-6327, 765-463-7200	www.mdbs.com

Anexo 2
Back Ends

Back-End	Proveedor	Dirección	URL
UniData	Ardent Software Inc.	Westboro, MA; 800-959-1221, 303-294-0800	www.ardentsoftware.com
Universal Data Option for Informix Dynamic Server	Informix Software Inc.	Menlo Park, CA;	www.informix.com
Velocis Database Server	Raima Corp.	Seattle, WA; 800-327-2462, 206-557-0200	

No. Prov	Proveedor	Dirección	Teléfono(s)	e-mail	URL	Fecha de creación
1	AKT Systems	P.O. Box 452, East Caulfield 3145, Victoria, Australia	613 9885 7171			
2	AbTech Corporation	Address.-1575 State Farm Blvd., Suites 1 & 2 Charlottesville, VA 22911	(804) 977-0686		www.abtech.com	1988
3	Advanced Software Applications Corp	333 Baldwin Road Pittsburgh, Pennsylvania 15205	412.429.1003, Fax: 412.429.0709	asa@asacorp.com	www.asacorp.com	Pittsburg, Pennsylvania 1992
4	Business Objects	Norte América: Business Objects Americas 2870 Zanker Road, EUROPA: Austria DELPHI GesmbH, Belgium Business Objects BeLux S.A./NV 6 Minnervastraat 1930 Zaventem.	America: +1 408 953 6000, +1 800 527 0580 Fax:+1 408 953 6001, Austria: +43 1 815 14 56 0 Fax:+43 1 815 14 56 21, Belgica: +32 2 720 0085 Fax:+32 2 720 7121	webmaster@businessobjects.com	www.businessobjects.com	1990
5	Cognos	France Tour Aurore 18, place des Reflets 92975 Paris La Défense 2 Cedex	33 (0)1 46 96 08 35, Fax: 33 (0)1 47 73 73 94	webmaster@cognos.com	www.cognos.com	1969
6	Data Distilleries	Kruislaan 419, 1098 VA Amsterdam, The Netherlands	31 20 560 8433, Fax: +31 20 668 5486	info@ddi.nl	www.ddi.nl	1995
7	Datasage	Datasage, Inc.19 New Crossing Road Reading, MA. 01867	(781) 942-3600, Fax: (781) 942-2163	webmaster@Datasage.com	www.datasage.com	1991
8	MIT - Management Intelligenter Technologien GmbH	Promenade 9, 52076 Aachen, Germany	+49 2408 94580		www.mitgmbh.de	1991

Anexo 3
Software para minería de datos

No. Prov.	Proveedor	Dirección	Teléfono(s)	e-mail	URL htt://	Fecha de creación
9	AcknoSoft	58 rue du Dessous-des-Berges, 75013 Paris, France US Office: 3119 Inwood Ct, Sugar Land, TX 77478	Francia: (331) 46 43 55 73, USA :(713) 952 80 76	marcominc@magic.fr	www.acknosoft.com/	Diciembre de 1991
10	Angoss	34 St. Patrick Street, Suite 200, Toronto, Ontario, Canada M5T 1V1	(416) 593-1122		www.angoss.com/	1983
11	Attar Software	Newlands Road, Leigh WN7 4HN, England; Two Deerfoot Trail On Partridge Hill, Harvard MA 01451, USA	44 (0)870 60 60 870, Fax: 44 (0)870 60 40 156; 508 456 3946, Free call 800 456 3966, Fax: 508 456 8383.	Info@attar.co.uk, info@attar.com	www.attar.com/	1985
12	Automatic Forecasting Systems, Inc	P.O. Box 563, Hatboro, PA 19040	(215) 675-0652, Fax: (215) 672-2534	autobox@icdc.com	darkstar.icdc.com/~autobox/	1975
13	AZMY Thinkware Inc	1450 Palisade Ave #M1D, Fort Lee, NJ 07024	Tel@201)947-188, Fax@201)947-1804	mail@azmy.com	www.azmy.com	1994
14	Cap Gemini bv	PO Box 2575, 3500 GN Utrecht, The Netherlands	+31 30 252 7044		www.capgemini.nl/	
15	Cogit	620 Folsom Street, Suite 200, San Francisco, CA 94107	(415) 908-1900		www.cogit.com	
16	Cognos	One Burlington Business, Center 67 South Bedford Street, Suite 200W, Burlington, MA, 01803-5164	(617) 229-6600		www.cognos.com	
17	Computer Science Innovations, Inc.	1235 Evans Road Melbourne, FL 32904	(407) 676-2923		www.csihq.com	1983
18	CrossZ Software Corporation	60 Charles Lindbergh Boulevard, Uniondale, New York 11553	(516) 228-8500, Fax: (516) 228-8584	danp@crossz.com	www.crossz.com	1989

No. Prov.	Proveedor	Dirección	Teléfono(s)	e-mail	URL htt://	Fecha de creación
19	Data Distilleries	Kruislaan 419, 1098 VA Amsterdam, The Netherlands	+31 20 560 8433		www.ddi.nl	
20	DataMind Corp.	Suite 1200, 2121 South El Camino Real, San Mateo, CA 94403	(415) 287-2000		www.datamind.com	
21	Datasage	9 New Crossing Road, Reading, MA 1867	(617) 942-3600		www.datasage.com	
22	DAZ Systems Inc	1829 Palm Avenue, Manhattan Beach, CA 90266	(310) 546-6670		www.dazsi.com	1995
23	HNC Software Inc	111 Pacifica, Third Floor, Irvine, California 92718-3304	714-753-8010, Fax: 714-753-8020	info@hnc.com	www.hnc.com	1986
24	Hyperparallel	282 Second Street, 3rd Floor, San Francisco, CA 94105	(415) 284-7000		www.hyperparallel.com	
25	IBM	IBM Global Business Solutions, Route 100, Somers, NY 10589			direct.boulder.ibm.com/bi/tech/mining/intminer.htm	
26	Information Discovery, Inc.	703-B Pier Avenue Suite 169, Hermosa Beach, CA 90254	(310) 937 3600		www.datamine inter.net/	
27	Integral Decisions Systems SE GmbH (INDECS)	Belfortstrasse 8, 81667 München, Germany	+49 89/4471760		www.indecs.com	
28	Integral Solutions Ltd.,	Berk House, Basing View, Basingstoke, Hampshire, RG21 4RG, England	+44125655899, Fax. +44125663467	clementine@isl.co.uk	www.isl.co.uk	mediados de 1993
29	Intrepid Systems	1301 Harbor Bay Parkway, Suite 200, Alameda, CA 94502-6576	(510) 769-4888		ww.intrepidsys.com	
30	ISID. Ingeniería y Sistemas de Información y Documentación.	Avda. de España, 74, Bajo B, 28220 - Majadahonda - Madrid	+34. 1. 634 65 44, Fax. +34. 1. 634 56 16	e-mail: info@isid.es	www.isid.es/home.htm	1996
31	Isoft	Chemin de Moulon, F91190 Gif sur Yvette, France	+33 1 69 41 27 77		www.alice.fr/	1987

Anexo 3
Software para mineria de datos

No. Prov.	Proveedor	Dirección	Teléfono(s)	e-mail	URL htt://	Fecha de creación
32	IVEE Development AB	Stora Badhusgatan 18-20, S-411 21 Göteborg Germany	+46 31 701 4260	info@sportfire.com	www.ivee.com/	1996
33	LEVEL\5	1335 Gateway Drive, Suite 2005, Melbourne, FL 32901	(407) 729-6004		www.l5r.com	1984
34	Logica	32 Hartwell Avenue, Lexington, MA 02173	(617) 476 8000		www.logica.com/products/product-Discovery.html	
35	Magnify, Inc.	100 South Wacker Drive, Chicago, IL 60606, Suite 1130,	(312) 214-1420, fax: (312) 214-1429	Info@magnify.com	www.magnify.com	
36	Mathsoft	MathSoft, Inc.1700, Westlake Ave., N., Suite 500, Seattle, WA 98109	(617) 577-1017	doug@statsci.com	www.mathsoft.com/splus/	
37	Megaputer Intelligence	Headquarters: B. Tatarskaia 38, Moscow 13184 Russia, US office: 1518 E Fairwood Drive, Bloomington, IN 47408	(812) 325-3026		www.megaputer.ru	1993
38	MIT – Management Intelligenter Technologien GmbH	Promenade 9, 52076 Aachen, Germany	+49 2408 94580		www.mitgmbh.de	1991
39	NCR	17095 Via del Campo, San Diego, CA 92127	(619) 485-3960		www.ncrhitc.com/high/skills/kd/	
40	NeoVista Solutions, Inc.	10710 North Tantau Ave.Cupertino, CA 95014	408-777-2929 , Fax: 408-777-2930	webmaster@neovista.com.	www.neovista.com	1981
41	NeuralWare, Inc.	202 Park West Drive, Pittsburgh, PA 15275	(412) 787-8222		www.neuralware.com	1987
42	Norsys Software Corp.	2315 Dunbar St., Vancouver, BC, Canada V6R 3N1			www.norsys.com	
43	Pilot Software	1 Canal Park, Cambridge, MA 02141	(617) 374-1000		www.pilotsw.com	
44	Prevision	1947 NW Garryanna St., Corvallis,	(541) 754-0569		www.prevision.com	
45	Promised Land	Promised Land Technologies, 195 Church Street 8th Floor, New Haven, CT 06516	(203) 562-7335		www.promland.com/index.html	

Anexo 3
Software para minería de datos

No. Prov.	Proveedor	Dirección	Teléfono(s)	e-mail	URL htt://	Fecha de creación
46	Quadstone Ltd	16 Chester Street, Edinburgh EH3, 7RA Scotland	+44 131 220 4491		ww.quadstone.co uk/	1994
47	Recognition Systems	Headquarters: Unit 6, Ashted Lock, Aston Science Park, Birmingham B7 4AZ UK, US Office: 10 South Riverside, Plaza, Suite 810, Chicago, IL 60606	(312) 382-8989			1989
48	REDUCT & Lobbe Technologies	P.O. Box 3570, Regina, Saskatchewan S4P 3L7, Canada	(306) 586-9400		ourworld.compuserve .com/homepages/red uct/	
49	Salford Systems	8880 Rio San Diego Dr., Suite 1045, San Diego, CA 92108	(619) 543-8880		www.salford-systems.com	1983
50	SAS	SAS Campus Drive, Cary, NC 27513-2414	(919) 677-8000		www.sas.com/feature /4qdm/intro.html	
51	Scholz & Theß Software GbR	Zwickauer Str 221, 09116, Chemnitz, Germany	+49 3 71 / 3 93 83 53		www.chemnitz-info.com/softwaregbr/	
52	Sentient Machine Research B.V.	Baarsjesweg 224, 1058 AA Amsterdam Holland	+31 20 6186927		www.xxlink.nl/smr/	1990
53	Siemens Nixdor	Siemens Nixdorf Advanced Technologies GmbH, Abteilung NC5, Rödelheimer Landstraße 5-9, D-60487 Frankfurt Germany	+49 69 797-4892		www.snat.de/nc5/	
54	Silicon Graphics S.A. de C.V.	Av. Vasco de Quiroga No. 3000, Col. Santa Fe, Cp. 01210, México D.F., 2011 N. Shoreline Blvd., Mountain View, CA 94043	Tel. (52-5) 267-1300 Fax (52-5) 267-1302	webmaster-mx@www.sgi.com	www.sgi.com/Products/software/MineSet/	1992 en México
55	slp InfoWare	200 West Madison Street, Suite 2150, Chicago - IL 60606	(312) 407-6580		www.slp-infoware.com/	
56	SPSS Inc.	444 N. Michigan Avenue, Chicago, Illinois 60611	(800) 543-2185		www.spss.com	
57	SRA International	444 N. Michigan Avenue, Chicago, Illinois 60611	(800) 543-2185			
58	Syllogic	Headquarters: Syllogic BV: De Molen 25, 3994 DA Houten, The Netherlands US Office: 6836 Leyland Park Drive, San Jose, CA 95120	(408) 268-4251		www.syllogic.nl	
59	Thinking Machines Corporation	14 Crosby Drive, Bedford, MA 01730	(617) 276-0400		www.think.com	1983

Anexo 3
Software para minería de datos

No. Prov.	Proveedor	Dirección	Teléfono(s)	e-mail	URL htt://	Fecha de creación
60	Trajecta	611 S. Congress, Suite 420, Austin, Texas 78704-1736	(800) 250-2242		www.trajecta.com	
61	TreeAge	1075 Main Street, Williamstown, MA 1267	(413) 458-0104		www.treeage.com	
62	Triada	315 E. Eisenhower Parkway, Suite 311, Ann Arbor, MI 48108	(313) 663-8622		www.triada.com/	
63	Ultimate Resources	13631 Queensbury, Houston, TX 77079	(713)b461-7734		www.hal-pc.org/~jpbrown/	1981
64	Unica	Lincoln North, Lincoln, MA 01773	(617) 259-5900		www.ultranet.com/~unica	1991
65	Urban Science	200 Renaissance Center, Detroit, MI 48243	(313) 259-9900			
66	WhiteCross Data Exploration	Waterside Park Cookham Road, Bracknell RG12 1RB, UK	'+44 1344 300770		www.contact.co.uk/whitecross/white1.htm	
67	WizSoft	WizSoft Inc., 3 Beit Hillel Street, Tel Aviv, 67017 Israel, US Office: WizSoft Inc., 6800, Jericho Tpke, Suite 120W Syosset, NY 11791	(516) 393-5841		www.wizsoft.com/	1983

No. Prov.	Software	Plataforma cliente	Plataforma servidor	Base de datos	Método
1	Data Mining Tool (DMT)	Win 95, Unix	Configuración Cliente/Servidor no soportado		Hibrido
2	ModelQuest Expert	Win 95, Win NT, Unix	Configuración Cliente/Servidor no soportado	ODBC	Hibrido(Basado en redes neuronales)
3	ScorXPRESS ®, ModelMAX ®, dbPROFILE ™, DecisionPOS ™	Win 95, Win NT	Configuración Cliente/Servidor no soportado	ODBC	Cluster, segmentación, Análisis descriptivo,
4	BusinessMiner	Win 95, Win NT, Win 3.1	Configuración Cliente/Servidor no soportado		Arboles de decisión, Análisis Si-entonces
5	Cognos Suite				
6	Data Surveyor 2.0		Unix		
7	Datasage				Multi-threaded parallel algorithm, SQL
8	DataEngine	Win 95, Win NT	Configuración Cliente/Servidor no soportado	ODBC	Redes neuronales y técnicas de lógica difusa.
9	KATE	Win 3.1, Win 95, Win NT	Configuración Cliente/Servidor no soportado		Arboles de decisión
10	KnowledgeSeeker	Win 3.1, Win 95, Win NT	Win NT, Unix	ODBC	CART
11	XpertRuler Profiler, XpertRuleKBS, XpertRule Configurator	Win 3.11, Win 95, Win NT	Configuración Cliente/Servidor no soportado	ODBC	Modelos de regresión en ID3

Anexo 3
Software para minería de datos

No. Prov.	Software	Plataforma cliente	Plataforma servidor	Base de datos	Método
12	AUTOBOX	RS6000's and DEC-ALPHA; todas las plataformas que soporten FORTRAN.			
13	SuperQuery	Win 3.x, Win 95, Win NT	Configuración Cliente/Servidor no soportado	ODBC, archivos planos, MS-Access, xBASE, Excel, Paradox	
14	Omega				
15	Knowledge Discovery Center				
16	Scenerio, 4Thought				
17	Visualizer WS	Win NT	Configuración Cliente/Servidor no soportado	ODBC	Normalizacion, extraccion , estadisticas, AutoCluster y sucesiones basadas en conocimiento (KBT).
18	QueryObject TM y Voyager	Win 95, Win NT, UNIX, MVS	Win NT, UNIX, MVS (HP, SUN, DEC, MS-NT)	ODBC	Algoritmos matematicos, estadística.
19	Data Surveyor				
20	DataMind				
21	Datasage				
22	AgentBase	Win 95, Win NT	Win NT, Unix	Archivos planos (los datos son replicados en formato propietario)	Híbrido (Redes neuronales, algoritmos genéticos, reglas de inducción)
23	DataBase Mining® Marksman				Redes neuronales.
24	Discovery				
25	Intelligent Miner				

No. Prov.	Software	Plataforma cliente	Plataforma servidor	Base de datos	Método
26	IDIS				Tecnología de lógica difusa, redes neuronales y estadística convencional.
27	DAFS (Beta in November 1997)				
28	Clementine	Win 95, Win NT, Unix, VMS	Win NT, Unix, VMS	Oracle, SQL Server, Informix, Ingres, Access, Teradata, Sybase	Reglas de inducción, redes neuronales, modelos de regresión y redes Kohonen, Visualización, estadísticas, reglas de inducción, redes neuronales, reglas de asociación.
29	DecisionMaster				
30	Spotfire Pro				Las tecnicas de Mineria de Datos Visual desarrolladas por
31	Alice, AC2	Win NT, Win 95, Win 3.x, UNIX (AC2 only)	Configuración Cliente/Servidor no soportado	ODBC, SAS, SPSS, BusinessObjects, Andyne GQL, Texto delimitado), Texto (de longitud fija), Access, Paradox, Dbase, Foxpro, Excel,Lotus, Btrieve	CART, ID3/C4.5, Técnicas avanzadas de visualización
32	Spotfire Pro	Win 95, Win NT, Unix	Configuración Cliente/Servidor no soportado	ODBC, archivos planos	Exploración interactiva y visualización de tendencias y patrones
33	Quest, Discovery 3D Toolkit	Win 95, Win NT	Win NT, Sun Solaris	ODBC	Hibrido (basado en redes neironales)
34	Discovery 3D Toolkit				

Anexo 3
Software para minería de datos

No. Prov.	Software	Plataforma cliente	Plataforma servidor	Base de datos	Método
35	PATTERN:Detect™, PATTERN:Profit™ , Herramientas: PATTERN:Classify™, PATTERN:Predict™, PATTERN:Optimize™, PATTERN:Store™		SUN, HP, IBM, y Tandem		CART
36	S-Plus, Trellis Graphics Displays: A Multi-Dimensional Data Visualization Tool for Data Mining				SQL
37	PolyAnalyst, PolyAnalyst Knowledge Server	Win95, WinNT, OS/2 Warp	WinNT, OS/2 Warp	ODBC, Oracle, Informix, DB2	Programación Evolutiva y Adquisiciones simbolicas del conocimiento
38	DataEngine	Win 95, Win NT	Configuración Cliente/Servidor no soportado	ODBC	Redes neuronales y técnicas de lógica difusa.
39	Management Discovery Tool (MDT)				
40	NeoVista's Decision Serie	Win 95, Win NT	HP, DEC, Sun SMP's	ODBC, Oracle, Informix, Sybase	Redes neuronales, árboles y reglas de inducción, Clustering, Asociación, reglas de extracción, Sequencing
41	NeuCOP, NeuralWorks Predict, NeuralWorks Professional II/PLU				Redes neuronales
42	Netica				
43	Discovery Server	Win 95, Win NT, Win 3.1	Win NT, Unix (HP, SUN,AIX) Informix, Sybase	Oracle, Microsoft SQL Server,	CART
44	Strategist, Preclass				
45	Braincel				Redes neuronales

Software para minería de datos

No. Prov.	Software	Plataforma cliente	Plataforma servidor	Base de datos	Método
46	Decisionhouse	UNIX (X Windows)	Sun, SGI, NCR, Digital flat files, Sybase	Oracle, Teradata, Non-stop SQL,	Arboles de decisión
47	Relationship Manager			ODBC	Redes neuronales
48	DATALOGIC/R				
49	C5 0/See5, Cubist	Win 95, Win NT, Unix	Ninguna		Clasificación (árboles de decisión, conjunto de reglas), Regresión (Basado en reglas, modelos lineales)
50	CART	Dos, Windows 3 1, 95 NT, Mac, UNIX		Archivos planos	CART
51	Enterprise Miner				
52	STARC				
53	DataDetective	Win 95, Win NT		ODBC, archivos planos	Propietario (basado en Nearest Neighbor)
54	SENN				Redes neuronales
55	MineSet	Silicon Graphics (corriendo con Irix 6.2 o superior), PCs (corriendo X-servers con OpenGL que puedan conectarse a la plataforma de Silicon Graphics)			Visualización, árboles de decisión, árboles de opción , naive-bayes/evidence, asociaciones
56	User Information System, Customer Profiling System				
57	SPSS Chaid				
58	KDD Toolset (Java tools)				
59	Syllogic Datamining Tool/MP				
60	Darwin				CART, Redes neuronales, Nearest Neighbor, y algoritmos genéticos
61	dbProphet				Redes neuronales

Anexo 3
Software para minería de datos

No. Prov.	Software	Plataforma cliente	Plataforma servidor	Base de datos	Método
62	DATA	Win 95, Win NT, Win 3.1, Macintosh			Arboles de decisión
63	Ngram				
64	SuperInduction	Win 95, Win NT		Texto, Dbase	Proprietario (basado en redes neuronales)
65	Model 1	Win 95, Win NT		ODBC	Hibrido
66	Gain			SAS data sets	
67	HeatSeeker	Win 3.1, Win 95	WX 9010, 9020	Propietario	CHAID
68	WizRule for Windows	Win 95, Win NT	Configuración Cliente/Servidor no soportado	ODBC, text	Reglas si-entonces

Anexo 3
Software para minería de datos

No. Prov.	Descripción	Clientes
2	Herramienta de modelado predictivo que ofrece las tecnologías de modelado de Redes Estadísticas y NEW StatNet Expert, puede importar y exportar datos de Excel, Access, Visual Basic, Paradox.	Universidades, y organizaciones de gobierno tales como: U.S. State Department, CitiBank, Exxon, University of Virginia Health Sciences Center, Mayo Clinic, DuPont, Dow Chemical, Moody's, and Oppenheimer Capital Investments.
3	Un conjunto integrado de herramientas para mineria d de datos automatizada, análisis de datos y soporte de decisiones para empresas grandes.	Empresas de mercadotecnia, Bancos
4	Es una poderosa herramientas de mineria de datos que permite encontrar la información escondida en los datos de una empresa que se encuentran en bases de datos relacionales, datawarehouses o en archivos individuales Descubre patrones en los datos oculto	
5	Es una solución completa para negocios inteligentes, es una herramienta poderosa para el análisis multidimencional, creación de escenarios para el descubrimiento de patrones y relaciones y adecuado para la generación de reportes búsquedas imprevistas.	Domino's Pizza, SutterHome Winery, York, Intrenational Business Goup
6	Es una interface gráfica de usuario(100% java) que permite: formular cuestionamientos de mineria, inspeccionar datos y resultados e interactuar con la guía del proceso de mineria de datos.	
7	Permite la explotación sistemática y la toma de decisiones automática. Requiere de la integración de los sistemas operaciones.	
9	Retoma las experiencias pasadas que son similares al problema que está ocurriendo y adapta la solución que se trabajó en el pasado al problema actual	Atlantis Aerospace (Canada), Compagnie Bancaire (France), French Ministry of Defense (France), VTT Electronics (Finland), Electricité de France (France), Daimler Benz (Germany), British Airways (UK)
10		CIBC, R.R. Donnelly, Ernst and Young

No. Prov.	Descripción	Clientes
31	Es un conjunto de librerías de C/C++ que permite implantar las funcionalidades del minado de datos, permite el modelado de datos.	Bank of America, Mc Kinsey, Mitre Corporation
32		3M, Unilever, Telia
33		American Express, SAIC, Marketing Management Inc.
35	Ofrece un sistema completo de minería de datos que integra un data warehouse, un conjunto de herramientas para minería de datos, un conjunto de herramientas para modelado predictivo	
36	Es una herramienta muy importante para la minería de datos, su objetivo es la visualización multidimencional de conjuntos de datos, presenta una gran variedad de tipos de gráficas o vista donde cada gráfica despliega la porción de datos seleccionados.	
37	Es una herramienta de minado de datos para usuarios expertos Esta compuesto por un conjunto de poderosos algoritmos que dan soporte a todo el proceso de descubrimiento de conocimiento	Moscow Energy Co., MI Bank, Ulm, University Clinic
40	Son un conjunto de herramientas que usan tecnicas avanzadas de descubrimiento de conocimiento que pueden ser aplazadas a cualquier tipo de empresa y a una gran variedad de problemas.	
41		DuPont, 3M, Texaco
46		Barclays Bank, Sainsbury's, Supermarket, British Airways
68	Descubre reglas e identifica excepciones en esas reglas, genera predicciones y clasificaciones	

Glosario

3GL: (Tercera Generación de Computadoras third-generation computer) Una computadora que usa circuitos integrados, almacenamiento de disco y terminales en línea. La tercera generación inició al rededor de 1964 con el sistema 360 de IBM.

ANSI: (American National Standards Institute). Instituto Nacional de Normas Americanas. Organización de establecimiento de normas patrocinadas por la industria. Fue fundada en 1918 y fijaba las normas industriales de los Estados Unidos y su correspondencia con las establecidas por al OSI (International Organization for Standarization). ANSI determina las normas relativas al hardware, en puntos tales como protocolos de nivel de enlace, posiciones y significado de las pastillas en los chips, registro en cinta y disco y algunas normas para el software.

API: (Application Program Interface) Interfaz de programador de aplicaciones. Especificación de la comunicación entre un programa de aplicaciones y uno de utilidad.

Asociaciones: Técnica utilizada en la minería de datos, el objetivo de la asociación es encontrar tendencias a través de un gran número de transacciones que pueden ser utilizadas para comprender y explorar los patrones naturales de compra, ajustar inventarios, analizar clientes, identificar servicios financieros que la gente compra

Banco de datos: Son un conjunto de registros listados uno después de otro, almacenados en archivos. Es una colección de datos que no tienen ningún tipo de relación entre ellos. (Definición propia)

Base de datos: Una base de datos es una colección de datos, que guardan una relación entre sí.

Clasificación: Técnica de análisis de la minería de datos, emplea un conjunto de transacciones preclasificadas para desarrollar un modelo que pueda clasificar los registros de una gran base de datos. Este tipo de análisis es particularmente adecuado para aplicaciones de análisis de riesgo y detección de fraudes. Emplea frecuentemente algoritmos basados en arboles o redes neuronales.

Cluster: Segmentación de problemas similares. Esta técnica es a menudo uno de los primeros pasos en el análisis de la minería de datos. Identifica grupos de relaciones entre los registros que pueden ser usadas como un punto de partida para futuras exploraciones en las relaciones; además, ayuda a determinar las características de los segmentos con respecto a los resultados deseados.

CODASYL: (COnference on DAta SYstems Languages) Formada por representantes de las áreas de la defensa y gobierno de los Estados Unidos, así como representantes mundiales de las áreas de negocios, con el objetivo inicial de proponer un lenguaje de programación de alto nivel para usar en el desarrollo de programas aplicables a negocios. Este objetivo inicial fue alcanzado con la publicación del primer COBOL.

Data warehouse: Almacén de datos, es una colección de datos integrales, variables en el tiempo, no volátiles y orientados a temas importantes para el soporte a la toma de decisiones en al administración de la organización.

Datawharehousing: Es el proceso de extracción y transformación de datos obtenidos en fuentes operacionales (OLTP) llevándolos a una base de datos centralizada reconocida como un data

warehouse Por lo que una vez en el repositorio es explotada usando herramientas para la toma de decisiones.

Dato: Del latín Datum, lo que se da. Antecedente necesario para llegar al conocimiento exacto de una cosa o para deducir las consecuencias legítimas de un hecho.
Se emplea el término dato para referirse a los valores registrados físicamente en la base de datos.

Datos extraños o irregulares: Son datos desconocidos o no comunes en la información que se esta manejando.

Deadlocks: Interbloqueo, bloqueo, enlazamiento fatal. Forma especifica de bloque (o abrazo fatal) que se presenta en una red, en la que algunos estados de la red se hacen inaccesibles para siempre.

Downsizing: Es mover las bases de datos corporativas de sistemas grandes y centralizados, a sistemas pequeños y menos costosos que no requieren de soporte y mantenimiento extensivos

DBMS: *(Data Base Management System)* Es una pieza sofisticada de software, la cual soporta la creación, manipulación y administración de sistemas de bases de datos. Esto significa que una base de datos y un DBMS, además de tener capacidades de propósito especifico, pueden ser usadas para reemplazar un DBMS de propósito general. Un sistema de bases de datos constituye en sí mismo un sistema completo de información; más comúnmente, es un componente de un gran sistema con otros componentes que incluyen programas que hacen uso de sus facilidades.

Encriptación: Proceso por el que un remitente convierte un mensaje inteligible para receptores a autorizados en ininteligible para otras personas no autorizadas.

Estándar: Una especificación para hardware o software que está extensamente usada y aceptada (de facto) o que esta sancionada por una organización de las normas (de jure).

FMS: *(Sistema Manejador de Archivos)* Es el único modelo que describe cómo son almacenados los datos en el disco. En este modelo, cada campo o dato es almacenado secuencialmente sobre el disco en un gran archivo. Fue el primer método usado para almacenar datos en una base de datos computarizada y la simplicidad es su única ventaja. Los productos existentes actualmente sobre este modelo son de bajo nivel. Sus desventajas son claras. Primero, no hay otra indicación de la relación entre los elementos más que la secuencia de almacenamiento. El programador, y algunas veces el usuario, tiene que conocer exactamente como son almacenados los datos en el archivo para poder manipularlos.

Host: Computadora unida a una red que proporciona servicios distintos de la simple actuación como procesador de almacenamiento y envío de conmutador de comunicación.

Información: Cualquier mensaje capaz de ser representado y manipulado. Resultado del tratamiento de los datos de un programa presentado en un formato determinado.

Interface: La conexión e interacción entre hardware, software y los usuarios. Frontera común entre dos sistemas, dispositivos o programas.

ISO: *(International Standards Organization)* Modelo de referencia ISO/OSI. Arquitectura general propuesta por la Organización Internacional de Normas para los Sistemas de Comunicación que permite la conexión de sistemas abiertos.

Join: En el manejo de bases de datos, combinar un archivo con otro de acuerdo a una condición determinada creando un tercer archivo con datos de los archivos comparados.

Mainframe: Computadora grande. Al principio todas las computadoras eran "grandes", puesto que era el adjetivo aplicado al gabinete que alojaba la unidad central de procesamiento: Este término se usa para denominar la parte básica o primordial de una computadora.

Mapear: Descripción de la forma en que se asocian entre sí, los registros de archivos, diferentes, en una base de datos.

MDD: Acrónimo de Multidimensional Database

Microcomputadora: Computadora pequeña con su unidad aritmética - lógica (ALU) y una unidad de control contenidas en un circuito integrado llamado micro-procesador.

Minicomputadoras: Computadora de programas almacenados, teniendo generalmente menos memoria y menor número de palabras que las máquinas mayores, es una computadora para multiusuarios, diseñada para hacer frente a las necesidades de las compañías pequeñas.

Nodo: Es un elemento de un dato que puede ser accesado por dos o más rutas.

OLAP: (OnLine Analytical Processing database) Base de datos diseñada para accesar a datos resumidos. Usando técnicas especiales de indexación, procesa consultas que pertenecen a grandes cantidades de datos más rápido que una tradicional base de datos.

OLTP: (OnLine Transaction Processing) Procesamiento de transacciones tal y como son recibida por la computadora. También llamado en línea o sistemas en tiempo real, los archivos maestros son modificados tan pronto como las transacciones son introducidas en las terminales o llegan sobre las líneas de comunicación.

Plataforma: Arquitectura hardware de un modelo particular o de una familia de computadoras. Es el estándar con el cual los desarrolladores escriben sus programas. El término también se refiere a un sistema operativo, el cual implica una arquitectura de hardware especial.

RDBM: *(Relational Data Base Management System)* El modelo relacional abandona el concepto de relaciones padre-hijo entre diferentes elementos de datos. Además, el dato es organizado en conjuntos lógicos matemáticos dentro de una estructura tabular. En un RDBM, cada campo se convierte en una columna dentro de una tabla, y cada registro se convierte en un renglón. Diferentes relaciones entre varias tablas, son definidas a través del uso de funciones matemáticas, tales como el *JOIN* y *UNION*.

Redes neuronales: Técnica de análisis de la minería de datos que construye modelos atendiendo los patrones en los datos que se analizan. Son modelos predictivos que basados en principios similares a aquellos que rigen el cerebro humano. En una red de nodos (neuronas), cada nodo recibe una entrada y envía una salida a los nodos subsecuentes basándose en lo que recibió como entrada. Una vez que la red neuronal ha sido validada, puede ayudar a analizar y predecir eventos a partir de entradas de datos nuevos.

Registro: Conjunto de campos relacionados que almacenan información acerca de un elemento.

Sistemas de visualización: Técnica d análisis de la minería de datos que permite hacer descubrimientos analizando los datos de manera gráfica con muchas variables, y después ver patrones y relaciones que sería muy difícil determinar mediante algoritmos de máquina, sin importar las capacidades de cómputo del sistema.

SQL: *(Structured Query Language)* Lenguaje usado para preguntar y procesar datos en bases de datos relacionales. Originalmente desarrollado por IBM para mainframes.

UNION: En bases de datos relacionales, es la unión de dos archivos.

UNIVAC I: (UNIVersal Automatic Computer) Primer computadora comercial exitosa, introducida en 1951 por Remington Rand.

BIBLIOGRAFÍA

[1] **ADAD RUBÉN, Careaga** Alfredo y Miguel Ángel **Medina,** Fundamentos de las estructuras de datos relacionales, MEGABYTE, México 1983, p. 57-59. 226 p.

[2] **CARDENAS, Alfonso F.,** Sistemas de administración de bases de datos, Limusa, México 1985.

[3] **Diccionario de la Lengua Española Real Academia Española**
España - Calpe, S.A. 19. Edición.

[4] **SALEMI,** Joe, PC Magazine Guide to Clien/Server Databases, Emeryville, California: Ziff-Davis, 263 p.

[5] **VIZCAÍNO** Carlos, "Soluciones Avanzadas, Tecnologías de Información y Estrategias de Negocios", Año 4, Número 34, 15 de Junio 96

[6] **Larouse 1 Conjugación, Sinónimos y Antónimos**
Ediciones Larouse, Editora de Periódicos, S.C.L., México 1992.

[7] http://www.pilotsw.com/r_and_t/whtpaper/datamine/dmfnd.htm

[8] **Glosario de términos y siglas Diccionario Inglés - Español SÁNCHEZ VAQUERO,** Antonio Luis **Joyanes Aguilar,** Mc GrawHill México 1985.ISBN 968-451-785-0

[9] http://rcr.csun.edu/ChrisJ/computers/sql/sqlch3.htm, **Relational Databases**

[10] **MARK** L. Gillenson, Introducción a las bases de datos, IBM System Research Institute, McGraw-Hill
ISBN 968-422-303-X , 1988

[10] http://ciir.cs.umass.edu/info/query_syntax_demoOld.html, **Query Formulation**

[11] http://users.demag.rwth-aachen.de/donald/Diversen/W3Encyc/_quer003.htm, **Query**

[12] **GONZÁLEZ BONILLA,** Marisol, Introducción a los sistemas de información ejecutiva y la evaluación de herramientas para su desarrollo, Soluciones Avanzadas, Núm. 35, 15 de julio 1996

[13] **HARJINDER** S. Gill y Prakash C. **RAO,** DATA WHAREHOUSING La integración para la mejor toma de decisiones, Prentice Hall Hispanoamericana S. A.,1996, pp. 182, 240

[14] http://www.wenet.net:81/~jtmalone/ClientServerDevelopmentTools.html

[15] http://www.dbmsmag.com/pcdbms.html