



Universidad de San Carlos de Guatemala

Facultad de Ingeniería

Escuela de Ingeniería en Ciencias y Sistemas

“DESARROLLO DEL LABORATORIO PRÁCTICO DE LOS CURSOS
“INTELIGENCIA DE NEGOCIOS 1 Y 2”, DE LA MAESTRÍA EN TECNOLOGÍA
DE LA INFORMACIÓN Y LA COMUNICACIÓN, EN SUSTITUCIÓN DE LA
CONTRAPARTE DE DOCENTES PROCEDENTES DE LA INDIA”

Melvin Ronaldo Díaz Marroquín

Asesorado por el Ing. Jorge Armín Mazariegos

Guatemala, noviembre de 2009.

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA



“DESARROLLO DEL LABORATORIO PRÁCTICO DE LOS CURSOS
“INTELIGENCIA DE NEGOCIOS 1 Y 2”, DE LA MAESTRÍA EN TECNOLOGÍA
DE LA INFORMACIÓN Y LA COMUNICACIÓN, EN SUSTITUCIÓN DE LA
CONTRAPARTE DE DOCENTES PROCEDENTES DE LA INDIA”

TRABAJO DE GRADUACIÓN

PRESENTADO A JUNTA DIRECTIVA
DE LA FACULTAD DE INGENIERÍA
POR

MELVIN RONALDO DÍAZ MARROQUÍN
ASESORADO POR EL ING. JORGE ARMÍN MAZARIEGOS
AL CONFERÍRSELE EL TÍTULO DE
INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, NOVIEMBRE DE 2009.

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA



NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. Murphy Olympo Paiz Recinos
VOCAL I	Inga. Glenda Patricia García Soria
VOCAL II	Inga. Alba Maritza Guerrero de López
VOCAL III	Ing. Miguel Ángel Dávila Calderón
VOCAL IV	Br. José Milton De León Bran
VOCAL V	Br. Isaac Sultán Mejía
SECRETARIA	Inga. Marcia Ivónne Véliz Vargas

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Sydney Alexander Samuel Milson
EXAMINADOR/A	Inga. Claudia Liceth Rojas Morales
EXAMINADOR/A	Ing. César Augusto Fernández Cáceres
EXAMINADOR/A	Ing. Manuel Fernando López Fernández
SECRETARIO/A	Ing. Pedro Antonio Aguilar Polanco

HONORABLE TRIBUNAL EXAMINADOR

Cumpliendo con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de Ejercicio Práctico Supervisado (EPS) titulado:

DESARROLLO DEL LABORATORIO PRÁCTICO DE LOS CURSOS “INTELIGENCIA DE NEGOCIOS 1 Y 2”, DE LA MAESTRÍA EN TECNOLOGÍA DE LA INFORMACIÓN Y LA COMUNICACIÓN, EN SUSTITUCIÓN DE LA CONTRAPARTE DE DOCENTES PROCEDENTES DE LA INDIA,

tema que me fuera asignado por la Dirección de la Escuela de Ingeniería en Ciencias y Sistemas, con fecha de agosto 2008.

Melvin Ronaldo Díaz Marroquín

Guatemala, 26 de junio de 2009.

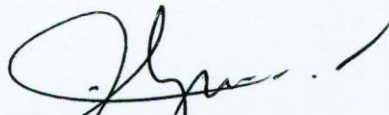
Ingeniera
Norma Ileana Sarmiento Zeceña de Serrano
Directora de la Unidad de EPS
Facultad de Ingeniería

Respetable Ingeniera Sarmiento de Serrano:

Por este medio hago de su conocimiento que he revisado el trabajo de graduación del estudiante MELVIN RONALDO DIAZ MARROQUIN, titulado "DESARROLLO DEL LABORATORIO PRÁCTICO DE LOS CURSOS "INTELIGENCIA DE NEGOCIOS 1 Y 2", DE LA MAESTRÍA EN TECNOLOGÍA DE LA INFORMACIÓN Y LA COMUNICACIÓN, EN SUSTITUCIÓN DE LA CONTRAPARTE DE DOCENTES PROCEDENTES DE LA INDIA", y a mi criterio el mismo cumple con los objetivos propuestos para su desarrollo, según el protocolo.

Sin otro particular, me suscribo de usted.

Atentamente,



Jorge Armín Mazariegos
Ingeniero en Ciencias y Sistemas
Colegiado No 5547
Asesor y Revisor de Trabajo de Graduación

E
S
C
U
E
L
A

D
E

C
I
E
N
C
I
A
S

Y

S
I
S
T
E
M
A
S


UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA



FACULTAD DE INGENIERÍA
ESCUELA DE CIENCIAS Y SISTEMAS
TEL.: 24767644

*El Director de la Escuela de Ingeniería en Ciencias y Sistemas de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor con el visto bueno del revisor y del Licenciado en Letras, de trabajo de graduación titulado **“DESARROLLO DEL LABORATORIO PRÁCTICO DE LOS CURSOS “INTELIGENCIA DE NEGOCIOS 1 Y 2”, DE LA MAESTRÍA EN TECNOLOGÍA DE LA INFORMACIÓN Y LA COMUNICACIÓN, EN SUSTITUCIÓN DE LA CONTRAPARTE DE DOCENTES PROCEDENTES DE LA INDIA”**, presentado por el estudiante MELVIN RONALDO DÍAZ MARROQUÍN, aprueba el presente trabajo y solicita la autorización del mismo.*

“ID Y ENSEÑAD A TODOS”


Ing. Marlon Antonio Pérez Turk

Director, Escuela de Ingeniería en Ciencias y Sistemas



Guatemala, 10 de noviembre 2009



UNIDAD DE E.P.S.

Guatemala, 04 de septiembre de 2009.
REF.EPS.DOC.1298.09.09.

Inga. Norma Ileana Sarmiento Zeceña de Serrano
Directora Unidad de EPS
Facultad de Ingeniería
Presente

Estimada Ingeniera Sarmiento Zeceña.

Por este medio atentamente le informo que como Supervisora de la Práctica del Ejercicio Profesional Supervisado, (E.P.S) del estudiante universitario de la Carrera de Ingeniería en Ciencias y Sistemas, **Melvin Ronaldo Díaz Marroquín** Carné No. **199312197** procedí a revisar el informe final, cuyo título es **“DESARROLLO DEL LABORATORIO PRÁCTICO DE LOS CURSOS “INTELIGENCIA DE NEGOCIOS 1 Y 2” DE LA MAESTRÍA EN TECNOLOGÍA DE LA INFORMACIÓN Y LA COMUNICACIÓN, EN SUSTITUCIÓN DE LA CONTRAPARTE DE DOCENTES PROCEDENTES DE LA INDIA”**.

En tal virtud, **LO DOY POR APROBADO**, solicitándole darle el trámite respectivo.

Sin otro particular, me es grato suscribirme.

Atentamente,

“Id y Enseñad a Todos”

Inga. Floriza Felipa Avila Pesquera de Medina

Supervisora de EPS

Área de Ingeniería en Ciencias y Sistemas

FFAPdM/RA





UNIDAD DE E.P.S.

Guatemala, 04 de septiembre de 2009.
REF.EPS.D.548.09.09.

Ing. Marlon Antonio Pérez Turck
Director Escuela de Ingeniería Ciencias y Sistemas
Facultad de Ingeniería
Presente

Estimado Ingeniero Perez Turck.

Por este medio atentamente le envío el informe final correspondiente a la práctica del Ejercicio Profesional Supervisado, (E.P.S) titulado **“DESARROLLO DEL LABORATORIO PRÁCTICO DE LOS CURSOS “INTELIGENCIA DE NEGOCIOS 1 Y 2” DE LA MAESTRÍA EN TECNOLOGÍA DE LA INFORMACIÓN Y LA COMUNICACIÓN, EN SUSTITUCIÓN DE LA CONTRAPARTE DE DOCENTES PROCEDENTES DE LA INDIA”**, que fue desarrollado por el estudiante universitario **Melvin Ronaldo Díaz Marroquín** Carné No. **199312197** quien fue debidamente asesorado por el Ing. Jorge Armin Mazariego y supervisado por la Inga. Floriza Felipa Ávila Pesquera de Medinilla

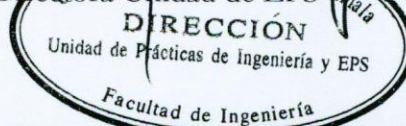
Por lo que habiendo cumplido con los objetivos y requisitos de ley del referido trabajo y existiendo la aprobación del mismo por parte del Asesor y de la Supervisora de EPS, en mi calidad de Directora apruebo su contenido solicitándole darle el trámite respectivo.

Sin otro particular, me es grato suscribirme.

Atentamente,

“Id y Enseñad a Todos”

Inga. Norma Ileana Serrano de Serrano
Directora Unidad de EPS



NISZ/ra



Universidad San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 14 de Octubre de 2009

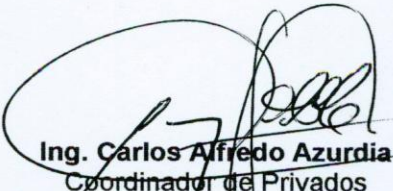
Ingeniero
Marlon Antonio Pérez Turk
Director de la Escuela de Ingeniería
En Ciencias y Sistemas

Respetable Ingeniero Pérez:

Por este medio hago de su conocimiento que he revisado el trabajo de graduación del estudiante **MELVIN RONALDO DIAZ MARROQUIN**, titulado: **“DESARROLLO DEL LABORATORIO PRACTICO DE LOS CURSOS INTELIGENCIA DE NEGOCIOS 1 y 2, DE LA MAESTRIA EN TECNOLOGIA DE LA INFORMACION Y LA COMUNICACIÓN, EN SUSTITUCION DE LA CONTRAPARTE DE DOCENTES PROCEDENTES DE LA INDIA”**, y a mi criterio el mismo cumple con los objetivos propuestos para su desarrollo, según el protocolo.

Al agradecer su atención a la presente, aprovecho la oportunidad para suscribirme,

Atentamente,


Ing. Carlos Alfredo Azurdia
Coordinador de Privados
y Revisión de Trabajos de Graduación





El Decano de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería en Ciencias y Sistemas, al trabajo de graduación titulado: **DESARROLLO DEL LABORATORIO PRÁCTICO DE LOS CURSOS "INTELIGENCIA DE NEGOCIOS 1 Y 2", DE LA MAESTRÍA EN TECNOLOGÍA DE LA INFORMACIÓN Y LA COMUNICACIÓN, EN SUSTITUCIÓN DE LA CONTRAPARTE DE DOCENTES DE LA INDIA**", presentado por el estudiante universitario **Melvin Ronaldo Díaz Marroquín**, autoriza la impresión del mismo.

IMPRÍMASE.

Ing. Murphy Olympo Paiz Rucinos
DECANO



Guatemala, noviembre de 2009

/cc
c.c. archivo.

ACTO QUE DEDICO A:

DIOS EL PADRE	Por su infinito amor hacia mí, enviando a su Hijo a dar su vida por mí.
JESUCRISTO	Mi salvador y mi redentor.
EL ESPÍRITU SANTO	Mi guía y mi consuelo, quien siempre está a mi lado.
MIS PADRES	Jorge Luis Díaz Hernández y Miriam Consuelo Marroquín Villeda, por su amor incondicional.
FAMILIA DIAZ VILLAGRAN	Mi querido hermano Aldo, la querida cuñis Amapola y mis sobrinos, Kathy, Kimberly y Aldo Javier.

AGRADECIMIENTOS A:

DIOS	Por su ayuda y bendiciones para alcanzar este punto tan importante para mi vida.
MIS PADRES	Por todo el esfuerzo y dedicación en formarme y convertirme en la persona que soy.
FAMILIA DIAZ VILLAGRAN	Por su apoyo y muestras de cariño en todo momento, y ser ese refugio cuando lo he necesitado.
MIS AMIGOS	Erick y Julio, por toda su amistad y apoyo en todo momento.
MIS HERMANOS	Alex, Geremías, por estar conmigo, sabiendo siempre qué decirme para cada situación.
COINSA Y COMPAÑEROS	Por la oportunidad que me ha dado de desarrollarme profesionalmente y crear grandes lazos de amistad y camaradería con todos los que allí laboramos.
MI ASESOR	Por toda la confianza y apoyo desinteresado que me brindó, para poder concluir con esta etapa.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES	III
GLOSARIO.....	VII
RESUMEN	IX
OBJETIVOS.....	XI
INTRODUCCIÓN	XIII
DESCRIPCIÓN DEL PROYECTO.....	XV
1. CONCEPTOS DE INTELIGENCIA DE NEGOCIOS.....	1
1.1. Datos, información y conocimiento.....	1
1.2. OLTP versus OLAP	2
1.3. <i>Datawarehouse</i> y <i>datamart</i>	4
1.4. Proceso de ETL	5
1.5. Persistencia MOLAP, ROLAP, HOLAP	6
1.6. <i>Datamining</i>	7
1.7. Por qué BI?	7
2. HERRAMIENTAS DE INTELIGENCIA DE NEGOCIOS	11
2.1. Microsoft® SQL Server® Analysis Services®	11
2.2. Microsoft® SQL Server® Integration Services®	12
2.3. Modelo Andamio.....	13

2.4. Tipificación de los ETL	14
3. DATAMINING.....	19
3.1. Descubrir el conocimiento (KDD).....	19
3.2. Técnicas para descubrir el conocimiento.....	19
3.3. Minería de datos	21
3.4. Algoritmos de minería de datos.....	22
3.5. Proceso de minería de datos.....	23
3.6. ¿Por qué minería de datos?	24
3.7. <i>Datamining</i> en la base de datos.....	25
3.8. Oracle [®] Data Mining (ODM).....	25
3.9. Funciones soportadas por ODM [®]	26
3.10. Ejemplos de aplicaciones de ODM	28
4. RESULTADOS.....	31
CONCLUSIONES.....	33
RECOMENDACIONES.....	35
BIBLIOGRAFÍA.....	37
APÉNDICE - MATERIAL ELABORADOR PARA LABORATORIOS.....	39

ÍNDICE DE ILUSTRACIONES

FIGURAS

1. Portada sesión 1, curso 1	39
2. Agenda sesión 1, curso 1.....	39
3. Instalación de SQL Server (R) 2005	40
4. Continuación instalación de SQL Server.....	40
5. Continuación de instalación	41
6. Repaso de Conceptos BI	41
7. Datos e información	42
8. Conocimiento	42
9. Pirámide de conocimiento	43
10. OLTP.....	43
11. OLAP	44
12. <i>Datawarehouse</i> y <i>datamart</i>	44
13. Características de <i>datawarehouse</i>	45
14. Portada sesión 2, curso 1	45
15. Agenda sesión 2, curso 1.....	46
16. ETL	46
17. Proceso de ETL	47
18. MOLAP, ROLAP y HOLAP	47
19. Datamining.....	48
20. Proceso de datamining.....	48

21. ¿Por qué BI?.....	49
22. Continuación ¿Por qué BI?	49
23. Bibliografía sesiones 1 y 2	50
24. Portada sesión 3, curso 1	50
25. Agenda sesión 3, curso 1	51
26. Componentes de SSAS.....	51
27. Continuación de componentes de SSAS.....	52
28. Continuación de componentes de SSAS.....	52
29. Usos típicos de SSIS	53
30. Ejemplos de tipos de ETL.....	53
31. ETL de extracción.....	54
32. ETL de inicialización	54
33. ETL de preparación	55
34. ETL de carga.....	55
35. ETL de procesamiento	56
36. Portada sesión 1 y 2, del curso 2	56
37. Agenda sesiones 1 y 2, curso 2.....	57
38. Repaso datos, información y conocimiento.....	57
39. Descubrir el conocimiento	58
40. Técnicas de KDD	58
41. Continuación técnicas de KDD	59
42. Minería de datos	59
43. División de la minería de datos	60
44. Aplicaciones de minería de datos	60
45. Técnicas de minería de datos.....	61
46. Algoritmos de minería de datos	61
47. Proceso de minería de datos.....	62

48. Extensiones de minería de datos	62
49. Por qué usar minería de datos	63
50. Portada sesión 3, curso 2	63
51. Agenda sesión 3, curso 2	64
52. Proceso de <i>datamining</i>	64
53. Continuación proceso de <i>datamining</i>	65
54. <i>Datamining</i> en BDD	65
55. Oracle (R) <i>Datamining</i>	66
56. Funciones de ODM	66
57. Continuación de funciones de ODM	67
58. Funciones soportadas por ODM	67
59. Continuación de funciones soportadas por ODM	68
60. Continuación de funciones soportadas por ODM	68
61. Portada sesión 4, curso 2	69
62. Agenda sesión 4, curso 2	69
63. Ejemplo 1 de ODM	70
64. Ejemplo 2 de ODM	70
65. Ejemplo 3 de ODM	71
66. Ejemplo 4 de ODM	71
67. Aplicación de ejemplos de ODM	72
68. Componentes de ODM	72

TABLAS

I - Resultados Laboratorio Inteligencia de Negocios 1	31
II - Resultado Laboratorio Inteligencia de Negocios 2	32

GLOSARIO

<i>Commit</i>	Palabra reservada en las bases de datos, utilizada para comprometer o grabar una transacción.
DMX	Acrónimo en inglés para expresiones de minería de datos (<i>DataMining eXpressions</i>).
DM	Acrónimo en inglés para minería de datos (<i>DataMining</i>).
Red bayesiana	“Una red bayesiana, o red de creencia, es un modelo probabilístico multi-variable que relaciona un conjunto de variables aleatorias mediante un grafo dirigido que indica explícitamente influencia causal”. ¹
MDX	Acrónimo en inglés para expresiones multidimensionales (<i>Multi Dimensional eXpressions</i>).

¹ http://es.wikipedia.org/wiki/Red_bayesiana consultado en diciembre 2008

<i>Rollback</i>	Palabra reservada en las bases de datos, utilizada para deshacer o reversar una transacción, y que los datos no se vean afectados.
SGBD	Sistema de Gestión de Base de Datos.
SQL	Acrónimo en inglés para Lenguaje estructurado de consulta (<i>Structured Query Language</i>).

RESUMEN

El presente trabajo se compone de una recolección de conceptos y definiciones relacionadas con la Inteligencia de Negocios, preparados para impartir el laboratorio de los cursos de Inteligencia de Negocios 1 y 2 de la Maestría en Tecnología de la Información y la Comunicación, de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala. Dicho trabajo se realizó con los estudiantes de la segunda promoción de dicha maestría.

Capítulo 1 – Inteligencia de Negocios, acá se describen los conceptos de lo que es la Inteligencia de Negocios, se repasa el concepto de la evolución de datos a información y luego a conocimiento, además se revisan algunas aplicaciones de la Inteligencia de Negocios, y se enumera una lista de razones, de por qué las empresas deben invertir en la Inteligencia de Negocios.

Capítulo 2 – Herramientas de Inteligencia de Negocios, en este capítulo se revisan los componentes de la arquitectura de una solución de Inteligencia de Negocios, tanto herramientas tecnológicas, como los componentes de Microsoft® SQL Server®, como conceptuales, como el modelo Andamio y las tipificaciones de los ETL.

Capítulo 3 – *Datamining*, en este capítulo se estudia el concepto de la minería de datos, se define el proceso de la minería de datos, donde se hace énfasis que la mayor parte del trabajo se consume en el procesamiento de los datos, y se enumeran algunas de las técnicas más usadas en la minería de datos. Adicionalmente se evalúa la herramienta ODM de Oracle ® y sus aplicaciones a la minería de datos.

Capítulo 4 – Material de Laboratorio, este capítulo contiene la impresión de todas las presentaciones utilizadas en el curso, para transmitir los conceptos y conocimientos a los estudiantes.

Al final del informe, se presentan los resultados obtenidos por los estudiantes en los laboratorios de ambos cursos, atendidos por el presente EPS, así como las conclusiones y recomendaciones

OBJETIVOS

- **GENERAL**

Recopilar una serie de definiciones relacionadas con la Inteligencia de Negocios, para que sirva de apoyo al laboratorio de los cursos Inteligencia de Negocios 1 y 2 de la Maestría en Tecnología de la Información y la Comunicación, de la Facultad de Ingeniería, de la Universidad de San Carlos de Guatemala.

- **ESPECÍFICOS:**

1. Impartir el laboratorio de los cursos Inteligencia de Negocios 1 y 2, en el cuarto trimestre del 2008 y primer trimestre del 2009, respectivamente.
2. Preparar material para futuras referencias para el laboratorio de éstos cursos.
3. Guiar a los estudiantes en la aplicación práctica de los conceptos aprendidos en el laboratorio de los cursos.

4. Evaluar el grado de conocimiento alcanzado por los estudiantes, y reportarlo al catedrático titular.

5. Transferir la experiencia adquirida a los estudiantes, para que ellos puedan poner en práctica de forma más inmediata sus conocimientos relacionados con Inteligencia de Negocios.

INTRODUCCIÓN

La Inteligencia de Negocios es un tema que recientemente ha tomado importancia en el mundo empresarial, ya que las empresas se han dado cuenta que la información es poder, y quien tenga la información correcta en el momento oportuno (conocimiento), es quien puede actuar acertadamente para sacar ventaja competitiva reaccionando rápidamente para adaptarse a los cambios en las tendencias.

Por esta razón, este tema ocupa dos cursos del pensum de estudios de la Maestría en Tecnología de la Información y la Comunicación, de la Facultad de Ingeniería, de la Universidad de San Carlos de Guatemala. El presente trabajo comprende una recolección de los principales conceptos relacionados con la Inteligencia de Negocios, así como el resumen de las presentaciones utilizadas en el laboratorio de los dos cursos, impartidos en la promoción 2008-2009 de dicha maestría.

En el laboratorio de estos dos cursos, se revisaron los principales conceptos relacionados con la Inteligencia de Negocios, y los estudiantes tuvieron la oportunidad de aplicar dichos conceptos utilizando dos de las principales herramientas existentes para realizar sus modelos, como son los productos de las empresas de software Microsoft® y Oracle®.

DESCRIPCIÓN DEL PROYECTO

El presente proyecto de Ejercicio Profesional Supervisado (EPS) consistió en recolectar información, elaborar presentaciones, realizar ejemplos, orientar y apoyar a los estudiantes en la elaboración de sus proyectos así como la calificación de los mismos.

En el curso Inteligencia de Negocios 1 se tuvo una reunión por semana con los estudiantes de la maestría, en las cuales se estudiaron los temas de repaso de conceptos de Inteligencia de Negocios, se instaló la herramienta Microsoft® SQL Server® Analysis Services (MSAS) versión 2005, se repasaron los conceptos de cubos, se les proporcionó una base de datos de ejemplo, y los estudiantes elaboraron un proyecto de Inteligencia de Negocios, de acuerdo al giro de negocio de los datos proporcionados.

En el curso Inteligencia de Negocios 2 se tuvo una reunión cada dos semanas, y en él se estudiaron los conceptos de descubrir el conocimiento, la minería de datos (*datamining*), las técnicas y algoritmos de la minería de datos, se instaló la herramienta Oracle® 10 g Enterprise Edition, con la opción de Oracle Data Mining, y la herramienta ODMiner, en esta ocasión los estudiantes fueron los encargados de obtener una base de datos, y sobre esta base de datos aplicar un proyecto de minería de datos, y dar sus conclusiones al respecto.

Para ambos cursos, la metodología utilizada fue, previo a las sesiones, investigación y elaboración de presentaciones y/o ejemplos, y durante la sesión la presentación participativa con los estudiantes, así como la realización de los ejemplos con ellos, adicional de las tareas de investigación que los estudiantes elaboraron.

1. CONCEPTOS DE INTELIGENCIA DE NEGOCIOS

1.1. Datos, información y conocimiento

Se puede definir la Inteligencia de Negocios (BI por sus siglas en inglés - *Business Intelligence*-) como el proceso de convertir los datos en información, y posteriormente la información en conocimiento.

1.1.1. Datos: son los elementos mínimos de información que por sí mismos son irrelevantes para la toma de decisiones. Ejemplo:

- 23,423.
- Juan Pérez.
- La Estrella.

1.1.2. Información: son datos procesados y que tienen significado, es decir que son datos que tienen un contexto y una utilidad. Ejemplo:

- Ventas del mes: 23,423 unidades.
- Vendedor: Juan Pérez.
- Marca: La Estrella.

1.1.3. Conocimiento: es la mezcla de la información y la experiencia (saber por qué y cómo) que se utiliza para adquirir nueva información, que a la vez ayuda a tomar decisiones acertadas, ejemplo:

- Ventas del mes 23,423, 15% más alto que el mes anterior.
- El vendedor Juan Pérez atiende a clientes mayoristas.
- La Estrella marca a la que se le incrementó 50% de fondos en publicidad, respecto del mes anterior.

1.2. OLTP versus OLAP

1.2.1. OLTP (*OnLine Transaction Processing*): el procesamiento de transacciones en línea, se caracteriza por:

- El acceso a los datos está optimizado para tareas frecuentes de lectura y escritura. Por medio de transacciones que se actualizan por medio del *commit* o el *rollback*.
- Los datos se estructuran según el nivel de la aplicación.

- Los formatos de los datos pueden variar entre un departamento y otro, debido a que tienen fuentes distintas.
- La historia de los datos normalmente se limita a los datos actuales o recientes.

1.2.2. OLAP (*OnLine Analytical Processing*): el procesamiento de análisis en línea se caracteriza por:

- Las bases de datos son especializadas para operaciones de lectura. La mayoría de operaciones que se realizan son las consultas, las inserciones, actualizaciones o eliminaciones son muy raras.
- Los datos están estructurados de acuerdo a las áreas de negocio, pero los datos se encuentra en un formato uniforme e integrado a lo largo de toda la organización.
- La historia de datos es a largo plazo, que puede ser de dos a cinco años, dependiendo del giro y la dinámica del negocio.

- Los datos que alimentan las bases de datos OLAP por lo general proceden de los sistemas operacionales existentes, por medio de procesos de extracción, transformación y carga (ETL).

1.3. *Datawarehouse y datamart*

1.3.1. *Datawarehouse*, su traducción literal es “bodega de datos”, y este término se utiliza definir a una base de datos corporativa, que integra y unifica la información de una o más fuentes de datos distintas, lo cual permite realizar análisis desde una gran variedad de perspectivas o puntos de vista y con grandes velocidades de respuesta.

1.3.2. *Datamart*, por lo general se identifica con *datamart* a una sección de un *datawarehouse*, especializado en un área de negocios específica, por ejemplo el *datamart* financiero, el *datamart* de ventas, el *datamart* de producción, pueden ser los componentes del *datawarehouse* corporativo de una empresa.

1.3.3. **Características:**

- **Integrado:** sin importar de dónde provengan los datos, en estas bases de datos se consolidan y se integran todos los datos de las distintas fuentes de la información, lo que los convierte en sistemas “centralizadores” de la información.

- **Temático:** se puede definir el *datawarehouse* como el sistema que contiene todos los temas de la empresa, y el *datamart* es el que contiene un “tema” específico o área de la empresa.
- **Histórico:** en estos sistemas se tiene la capacidad de manejar grandes volúmenes de datos, lo que permite tener mucho tiempo de historia almacenada, lo cual permite realizar análisis de comportamientos y tendencias.
- **No volátil:** una vez que la información es cargada en el *datawarehouse* ya no se elimina ni se modifica, ya que los sistemas están optimizados para almacenar la información por mucho tiempo.

1.4. Proceso de ETL

Es el proceso por medio del cual la información es trasladada desde su fuente original al *datawarehouse* y contiene procesos de extracción, transformación y carga (ETL - *Extraction, Transformation and Loading*) por medio de los cuales los datos de un sistema OLTP son trasladados a un sistema OLAP, con el fin de transformar los datos en conocimiento.

1.4.1. Extracción: procesos encargados de obtener la información de las distintas fuentes, las cuales pueden ser tanto internas como externas.

1.4.2. Transformación: es el proceso por medio del cual se limpia, depura, filtra, homogeniza y agrupa la información de las distintas fuentes.

1.4.3. Carga: consiste en la organización y actualización de los nuevos datos ya transformados y los metadatos en la base de datos.

1.5. Persistencia MOLAP, ROLAP, HOLAP

Dentro de un *datawarehouse*, la información se puede almacenar de cualquiera de estas formas:

1.5.1. MOLAP – OLAP Multidimensional, calcula agregaciones y estructuras en motores multidimensionales, tiene las combinaciones pre-calculadas, requiere mucho espacio.

1.5.2. ROLAP – OLAP Relacional, se basa en un motor de base de datos relacional, y realiza los cálculos de las agregaciones en el momento que se solicitan.

1.5.3. HOLAP – OLAP Híbrido, una combinación entre los dos esquemas anteriores, en el cual las estructuras se almacenan con la técnica de MOLAP y las agregaciones por medio de ROLAP.

1.6. *Datamining*

El ***datamining*** (*minería de datos*), se refiere al proceso que puede ser automático o semiautomático, por medio del cual se buscan patrones o tendencias dentro de la información (que pueden ser grandes volúmenes de datos almacenados en una base de datos), con el fin de explicar los datos por sí mismos.

1.7. Por qué BI?

La principal justificación de por qué es necesario un proyecto de Inteligencia de Negocios, se puede definir como la respuesta a las siguientes preguntas:

- “Observar ¿qué está ocurriendo?”²
- “Comprender ¿por qué ocurre?”³

² http://www.sinnexus.com/business_intelligence/index.aspx consultada en Agosto 2008

³ Idem 2

- “Predecir ¿qué ocurriría si...?”⁴
- “Colaborar ¿qué debería hacer el equipo?”⁵
- “Decidir ¿qué camino se debe seguir?”⁶

1.7.1. BI como una solución tecnológica

- Centralizar, depurar y afianzar datos
- Descubrir información no evidente para las aplicaciones actuales
- Optimizar el rendimiento de los sistemas

1.7.2. BI como una ventaja competitiva

- Seguimiento real del plan estratégico
- Aprender de errores pasados
- Mejorar la competitividad

⁴ http://www.sinnexus.com/business_intelligence/index.aspx consultada en Agosto 2008

⁵ Idem 4

⁶ Idem 4

- Obtener el verdadero valor de las aplicaciones de gestión

2. HERRAMIENTAS DE INTELIGENCIA DE NEGOCIOS

2.1. Microsoft® SQL Server® Analysis Services®

Esta herramienta es el motor de base de datos o SGBD de tipo multidimensional de Microsoft® SQL Server® y su propuesta para la administración de bases de datos tipo OLAP, los principales componentes de una base de datos de este tipo, de acuerdo a esta herramienta son los siguientes:

- 2.1.1. **Orígenes de datos:** establece los mecanismos de conexión hacia el origen relacional de los datos.
- 2.1.2. **Vistas de origen de datos:** describe una vista de origen de datos, un esquema relacional y consultas asociadas utilizadas para modelar uno o más orígenes de datos, en Analysis Services®.
- 2.1.3. **Cubos:** describe cubos y objetos de cubo, lo que incluye medidas, grupos de medida, relaciones de uso de dimensiones, cálculos, indicadores clave de rendimiento, acciones, traducciones, particiones y perspectivas.

- 2.1.4. **Dimensiones:** describe dimensiones, tipos de dimensión, almacenamiento de dimensiones y objetos de dimensiones, incluidos atributos, relaciones de atributos, jerarquías, niveles y miembros.

- 2.1.5. **Estructuras de minería de datos:** describe estructuras de minería de datos y objetos de minería.

- 2.1.6. **Funciones:** mecanismo de seguridad para controlar el acceso a los objetos.

- 2.1.7. **Ensamblados:** describe un ensamblado, una colección de funciones definidas por el usuario utilizadas para ampliar los lenguajes MDX y DMX.

2.2. Microsoft ® SQL Server ® Integration Services ®

Para las tareas de extracción, transformación y carga (ETL), ésta la herramienta que presenta Microsoft ® para agilizar y optimizar estos procesos, las tareas comunes que nos permite realizar son:

- 2.2.1. Mezclar datos de almacenes de datos heterogéneos.
- 2.2.2. Llenar bodegas y almacenes de datos (*datawarehouse* y *datamart*)
- 2.2.3. Limpiar y normalizar datos
- 2.2.4. Generar BI en un proceso de transformación de datos.
- 2.2.5. Automatizar las funciones administrativas y la carga de datos

2.3. Modelo Andamio

El modelo Andamio es una base de datos genérica que se utiliza para ayudar a construir un *datawarehouse* pero que solamente sirve como un puente entre el modelo OLTP y el modelo OLAP, se puede decir que es una base de datos “relacional de paso”, ya que nos permite relacionar los datos de un modelo al otro, por medio de relaciones de identificadores, y de esta forma mantener la trazabilidad entre un sistema y el otro, y de paso, ya que nos puede ayudar para realizar las tareas de conversión, depuración, consolidación y validación de los datos, para no afectar directamente el sistema OLTP, y antes de convertir los datos a un OLAP.

Su estructura es muy parecida a la estructura del sistema OLTP, con cierta información adicional, y que nos puede ayudar también, para consolidar la información de varios sistemas heterogéneos, ya que el objetivo es que sea un unificador de los datos, provenientes de distintas fuentes, y ésta base de datos será la encargada de procesar los distintos procesos de ETL.

2.4. Tipificación de los ETL

Los ETL son los procesos encargados de mover la información de un sistema a otro, en este caso, para mover la información desde el sistema OLTP hacia el sistema OLAP, se pueden clasificar en 5 categorías principales:

2.4.1. **Extracción:** obtienen la información del OLTP y la transfieren al Andamio.

- En el modelo Andamio existe forma de relacionar el registro con el registro equivalente en el OLTP.
- Para las tablas de tipo dimensión: el proceso se puede resumir como: Revisar existencia de registro en Andamio, si es nuevo lo transfiere, si es distinto actualiza y marca el registro, si está igual, lo deja como tal.

- Para las tablas de hechos: obtiene datos en rango de tiempo según la periodicidad definida, convirtiendo los identificadores del sistema OLTP a los identificadores del Andamio.

2.4.2. **Inicialización:** obtienen y cargan las tablas de dimensiones del Andamio hacia la estrella.

- Estos procesos aplican para las tablas de dimensiones
- Realizan procesos de consolidación (“aplanan” tablas)
- Filtran los registros que han sido modificados.
- Actualizan la información, proveniente del OLTP en la Estrella
- Crean valores “Default” para posibles valores nulos.

2.4.3. **Preparación:** sincronizan las llaves de la Estrella hacia el Andamio y hacen las validaciones.

- Se realizan sobre los hechos.

- Realizan pre-cálculos necesarios para la estrella, (conversiones)
- Obtienen el identificador de la estrella, para cada uno de los campos asociados, en base al id del andamio.
- Verifican la ausencia de nulos
- Se realizan validaciones iniciales, cantidad de registros, suma de montos, razonabilidad de las conversiones, etc.

2.4.4. **Carga:** mueven la información del Andamio hacia la Estrella.

- Una vez que los datos en los hechos están preparados y validados.
- Traslada la información desde el Andamio hacia la Estrella.
- Eliminan la información trasladada del Andamio, para estar listos para la próxima carga.

- Si se permite volver a cargar ciertos datos, deben ser capaces de actualizar esa información en la Estrella (más óptimo: borrar y volver a trasladar)

2.4.5. **Procesamiento:** actualizan la información de los cubos.

- Son los encargados de refrescar la información, ya actualizada en la Estrella (DW) hacia los cubos.
- Refrescan las dimensiones, si hubo cambios drásticos, re-construirlas, sino actualización incremental.
- Procesar cubos y modelos de minería de datos, puede ser incremental, refrescamiento o re-construcción.

3. DATAMINING

3.1. Descubrir el conocimiento (KDD)

El proceso de descubrir el conocimiento (KDD por sus siglas en inglés *Knowledge discovery*) se puede definir como la obtención de información potencialmente útil que se encuentra implícita en los datos, pero que nos es desconocida.

“El proceso de descubrir el conocimiento toma los datos tal como vienen, los transforma en información útil y entendible, procesando grandes cantidades de datos crudos, identificando los patrones significativos y relevantes y los presentan como conocimiento apropiado para satisfacer las metas del usuario.”⁷

3.2. Técnicas para descubrir el conocimiento

3.2.1. Método de clasificación: es el más usado, “agrupa los datos de acuerdo a similitudes o clases”⁸.

⁷ <http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm>
Consultado en enero 2009

⁸ Idem 7

3.2.2. Método probabilístico: utilizando modelos de representación gráfica, se basa en las probabilidades e independencia de los datos, “puede usarse en sistemas de diagnóstico, planeación y sistemas de control”⁹.

3.2.3. Método estadístico: “usa la regla del descubrimiento y se basa en las relaciones de los datos”¹⁰, “usado para generalizar los modelos en los datos y construir las reglas de los modelos nombrados”¹¹.

3.2.4. Método Bayesiano: es un modelo gráfico, usando “frecuentemente las redes Bayesianas, cuando la incertidumbre se asocia con un resultado que puede expresarse en términos de probabilidad, usado en sistemas de diagnóstico”¹².

⁹ <http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MonografiaMD.PDF> consultado en enero 2009.

¹⁰ Idem 9

¹¹ <http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm> Consultado en enero 2009

¹² Idem 11

3.3. Minería de datos

El *Datamining* o minería de datos, “es un mecanismo de explotación, consistente en la búsqueda de información valiosa en grandes volúmenes de datos. Está muy ligada a los *Datawarehouse* ya que realiza el análisis de archivos y bitácoras de transacciones, trabajando a nivel del conocimiento con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones”¹³.

La minería de datos, utiliza técnicas estadísticas para convertirse en una minería de datos predictiva, o bien técnicas de inteligencia artificial para realizar una minería de datos para el descubrimiento del conocimiento.

Sus principales aplicaciones son:

- Aspectos climatológicos
- Medicina
- Mercadotecnia
- Inversión en casa de bolsa y banca
- Detección de fraudes y comportamientos inusuales

¹³ <http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm>
Consultado en enero 2009

- Análisis de canasta de mercado
- Determinación de niveles de audiencia
- Industria y Manufactura

3.4. Algoritmos de minería de datos

3.4.1. Supervisados o predictivos: “predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos”¹⁴. Requieren la especificación de un objetivo, el cual puede tener atributos binarios (compra, no compra) o bien una lista de alternativas (color de sweater, rangos de salarios, etc.)

3.4.2. No supervisados o del descubrimiento del conocimiento: con estos algoritmos se “descubren patrones y tendencias en los datos actuales. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio de ellas”¹⁵. Se usan para encontrar estructuras intrínsecas, relaciones o afinidades en los datos, no tienen un objetivo específico, y pueden ser usados para encontrar agrupaciones naturales en los datos.

¹⁴ <http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm>
Consultado en enero 2009

¹⁵ Idem 14

3.5. Proceso de minería de datos

3.5.1. “Determinación de los objetivos: delimitar los objetivos que el cliente desea”¹⁶.

3.5.2. Pre procesamiento de los datos: “se refiere a la selección, limpieza, enriquecimiento, reducción y transformación de las bases de datos”¹⁷. En la vida real, los datos por lo general se encuentran “sucios”, con datos incorrectos o ausentes, éstos se deben “limpiar” antes de utilizarlos, filtrando, normalizando, tomando muestras, transformando en varias direcciones. Cerca del 80% del esfuerzo en un proyecto de DM es invertido en la preparación de los datos, ya que éstos son la entrada para algún algoritmo de minería de datos.

3.5.3. Determinación del modelo: análisis estadístico de los datos y visualización gráfica de los mismos como una primera aproximación.

3.5.4. Análisis de los resultados: verificar la coherencia de los resultados obtenidos y compararlos con los resultados estadísticos y gráficos

¹⁶ <http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm>
Consultado en enero 2009

¹⁷ Idem 16

3.6. ¿Por qué minería de datos?

- 3.6.1. “Contribuye a la toma de decisiones tácticas y estratégicas”¹⁸.
- 3.6.2. “Proporciona poder de decisión a los usuarios del negocio, y es capaz de medir las acciones y resultados de la mejor forma”¹⁹.
- 3.6.3. Genera modelos descriptivos que permiten a las empresas explorar y “comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales”²⁰.
- 3.6.4. “Genera modelos predictivos que permiten que relaciones no descubiertas a través del proceso del DM sean expresadas como reglas de negocio”²¹.

¹⁸ http://www.at-systems.es/soluciones/data_mining.htm consultado en enero 2009.

¹⁹ [http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_\(Vallejos\).pdf](http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_(Vallejos).pdf) consultado en enero 2009

²⁰ <http://www.monografias.com/trabajos26/data-mining/data-mining2.shtml> consultado en enero 2009

²¹ <http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm> Consultado en enero 2009

3.7. *Datamining* en la base de datos

Para el procesamiento de un modelo de minería de datos, éstos deben ser contruidos, probados, validados, administrados e implementados en un ambiente apropiado, ya que los resultados pueden ser procesados posteriormente como parte de cálculos específicos y por lo tanto se vuelve necesario almacenarlos en una base de datos permanente, este proceso puede involucrar la transferencia de información entre servidores, repositorios de datos, aplicaciones y herramientas, conversiones de formatos, etc.

Eliminando o reduciendo estos obstáculos, se puede ejecutar el proceso de minería de datos con mayor frecuencia, utilizando datos más actualizados, reduciendo el movimiento de los datos, lo cual se traduce el tiempo total del procesamiento de la minería de datos, y si los datos no abandonan la base de datos, se mantiene la seguridad sobre los mismos.

3.8. Oracle ® Data Mining (ODM)

Esta herramienta, integra la minería de datos, dentro de la base de datos Oracle ®. Debido a que los algoritmos de minería de datos operan nativamente sobre las tablas o vistas relacionadas, se eliminan los procesos de ETL por medio de una herramienta especializada.

Con el ODM, las tareas de minería de datos pueden ejecutarse asíncronamente e independientes de alguna interface como parte de una base de datos normal. Y las herramientas pueden ejecutarse en línea con comandos de Java o bien con PL/SQL.

ODM automatiza la mecánica de construcción, pruebas y aplicación de modelos, de manera que los esfuerzos se enfoquen en los aspectos de negocios del problema y no de detalles estadísticos y matemáticos.

3.9. Funciones soportadas por ODM®

3.9.1. Funciones supervisadas

3.9.1.1. **Clasificación:** agrupa los ítems en clases discretas y predice a qué clase pertenece un ítem.

3.9.1.2. **Regresión:** la aproximación y la previsión de valores continuos.

3.9.1.3. **Importancia del atributo:** identificar los atributos que son más importantes en la predicción de resultados.

- 3.9.1.4. **Detección de anomalías:** identificar los elementos que no cumplan las características de los datos "normales".
- 3.9.1.5. **Árboles de decisión:** una manera rápida y escalable de extraer información predictiva y descriptiva de una tabla de la base de datos, respecto de un objetivo especificado por el usuario; los árboles de decisión proveen reglas de fácil entendimiento.
- 3.9.1.6. **Aprendizaje activo:** un algoritmo mejorado de vector de máquina de soporte (una clase de detección de anomalías) que soporta grandes volúmenes de datos

3.9.2. Funciones no supervisadas

- 3.9.2.1. **Agrupación (*Clustering*):** encontrar agrupaciones naturales en los datos.
- 3.9.2.2. **Asociación modelos:** análisis de "canasta de mercado".
- 3.9.2.3. **Extracción de características:** creación de nuevos atributos (características) como una combinación de los atributos.

3.10. Ejemplos de aplicaciones de ODM

3.10.1. Problema: un vendedor al detalle desea incrementar sus ingresos, identificando sus principales clientes potenciales, para crear incentivos para ellos. También desea una guía en su almacenamiento, determinando los productos que más frecuentemente se compran juntos.

Solución: un modelo de clasificación se puede construir para determinar los clientes que están “dispuestos” más de un 75% de gastar más de US\$1000 el próximo año. Un modelo de Reglas de Asociación para crear un análisis de canasta.

3.10.2. Problema: una agencia gubernamental desea métodos más rápidos y confiables para identificación de posibles actividades fraudulentas para futuras investigaciones.

Solución: crear modelos de clasificación, *Clustering* y modelos de detección de anomalías para marcar los casos “sospechosos”.

3.10.3. Problema: un investigador bioquímico desea trabajar con miles de atributos asociados con una investigación de la efectividad de una droga.

Solución: una función de importancia de atributos para reducir el número de factores a un subconjunto manejable de atributos.

3.10.4. Problema: una compañía hipotecaria desea incrementar los ingresos, reduciendo el tiempo requerido para la aprobación de los préstamos.

Solución: un modelo de regresión puede predecir el mejor valor para una casa, eliminando la necesidad de una inspección en sitio.

4. RESULTADOS

En el curso Inteligencia de Negocios 1, al inicio se contaba con la presencia de seis estudiantes, al final solamente terminaron tres estudiantes, los cuales obtuvieron los resultados mostrados a continuación. En la tabla I se indica además cómo estaba compuesta la ponderación de las nota del laboratorio.

Tabla I - Resultados Laboratorio Inteligencia de Negocios 1

	Asistencia (10%)	Tareas (5%)	Fase 1 (15%)	Fase 2 (15%)	Fase 3 (15%)	Fase 4 (40%)	Nota Final (100%)
Estudiante 1	10	4	10	15	13	38	90
Estudiante 2	10	3	13	5	10	30	71
Estudiante 3	9	2	14	13	13	35	86
Promedio	9.7	3.0	12.3	11.0	12.0	34.3	82.3

En el curso Inteligencia de Negocios 2, los estudiantes que participaron fueron los mismos tres que concluyeron el curso anterior, la modalidad consistió en una clase bi-semanal, siempre tomándose en cuenta la asistencia, tareas de laboratorio y el desarrollo de un proyecto de curso, y los resultados así como la ponderación obtenida en los mismos, se detalla en la tabla II.

Tabla II - Resultado Laboratorio Inteligencia de Negocios 2

	Asistencia (15%)	Tareas (25%)	Proyecto (60%)	Nota Final (100%)
Estudiante 1	13	23	50	86
Estudiante 2	14	18	55	87
Estudiante 3	12	15	55	82
Promedio	13.0	18.7	53.3	85.0

CONCLUSIONES

1. La aplicación práctica de los conocimientos adquiridos en el curso, por medio de un curso adicional de laboratorio, ayuda a los estudiantes a afianzar los mismos.
2. A pesar de que en la actualidad existe mucha información relacionada con la Inteligencia de Negocios, es importante tener una fuente de referencia rápida y resumida.
3. Cuando los estudiantes inician la aplicación práctica de los conceptos adquiridos, muchas veces es necesaria la guía de alguien que ya haya recorrido ese camino, para tener una mejor comprensión de lo que está haciendo.
4. La ponderación y medición de los conocimientos adquiridos por los estudiantes ayuda a identificar la eficiencia en la transmisión del conocimiento.
5. El desarrollo de un proyecto de curso, ayuda a que los estudiantes puedan asimilar y aplicar de mejor manera el conocimiento adquirido en el mismo.

RECOMENDACIONES

1. Promocionar más la existencia de esta maestría, ya que tener solamente tres estudiantes en la segunda promoción es muy poco, lo cual puede provocar que se cierre tan importante postgrado.
2. Documentar más a los estudiantes acerca de los requisitos que se tienen para estudiar este postgrado, ya que los estudiantes que desertaron del curso, son profesionales que provienen de una carrera distinta a la Ingeniería en Ciencias y Sistemas, ya que esta maestría requiere sólidos fundamentos en tecnologías de información.
3. Alentar a los estudiantes a que puedan disponer de equipo móvil, ya que para ir desarrollando sus proyectos, se les dificulta mucho estar movilizándolo sus bases de datos y/o equipos.
4. Formar una biblioteca de bases de datos extensas y completas, con información, para que los estudiantes puedan realizar sus pruebas de minería de datos, ya que en la elaboración del proyecto relacionado con ese tema, el mayor inconveniente que tuvieron los estudiantes fue el obtener una base de datos que les ayudará a cumplir los objetivos del proyecto de *datamining*.

5. Organizar más conferencias, en donde los estudiantes puedan conocer las opciones comerciales de aplicaciones disponibles en el mercado, adicionales a las que utilizan en sus prácticas, para que puedan tener un mayor panorama de las distintas formas en que esas herramientas aplican los conceptos.

BIBLIOGRAFÍA

1. es.wikipedia.org/wiki/Datamart (Consultado 20-ago-2008)
2. es.wikipedia.org/wiki/Datawarehouse (Consultado 20-ago-2008)
3. es.wikipedia.org/wiki/ERP (Consultado 14-ago-2008)
4. es.wikipedia.org/wiki/ETL (Consultado 14-ago-2008)
5. es.wikipedia.org/wiki/OLAP (Consultado 20-ago-2008)
6. es.wikipedia.org/wiki/OLTP (Consultado 20-ago-2008)
7. Haberstroh, Robert. Oracle Data Mining Tutorial for Oracle Data Mining 10g Release 2, Oracle Data Mining 11g Release 1. Oracle USA, 2008.
8. Microsoft SQL Server 2005, Libros en pantalla, 2005.
9. [msdn.microsoft.com/es-es/library/ms175609\(SQL.90\).aspx](http://msdn.microsoft.com/es-es/library/ms175609(SQL.90).aspx) páginas similares (Consultado 06-sep-2008)
10. sinnexus.com/business_intelligence/index.aspx páginas similares (Consultado 14-ago-2008)
11. Taft, Margaret; Krishnan, Ramkumar; Hornick, Mark; Muhkin, Denis; Tang, George; Thomas, Shiby; Stengard, Peter. Oracle *Datamining* Concepts, 10g Release 2 (10.2). Oracle USA, 2005.

APÉNDICE - MATERIAL ELABORADO PARA LABORATORIOS

Figura 1 - Portada sesión 1, curso 1



Figura 2 - Agenda sesión 1, curso 1

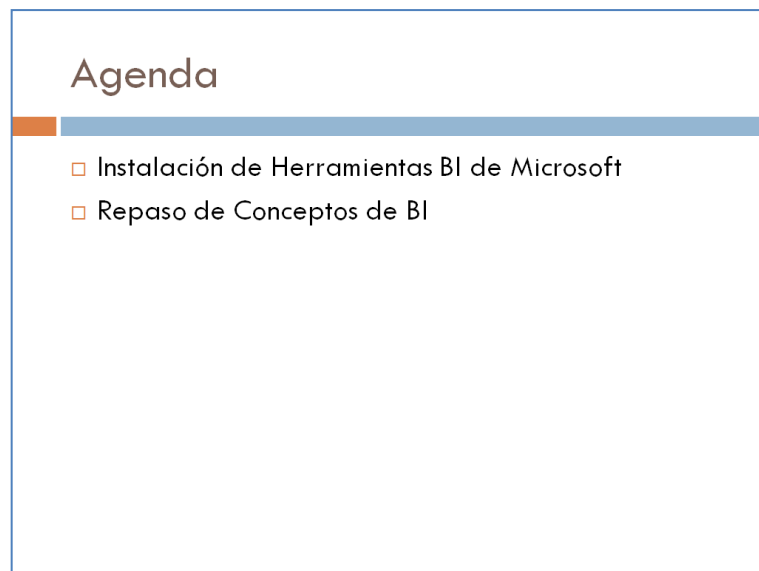


Figura 3 - Instalación de SQL Server (R) 2005

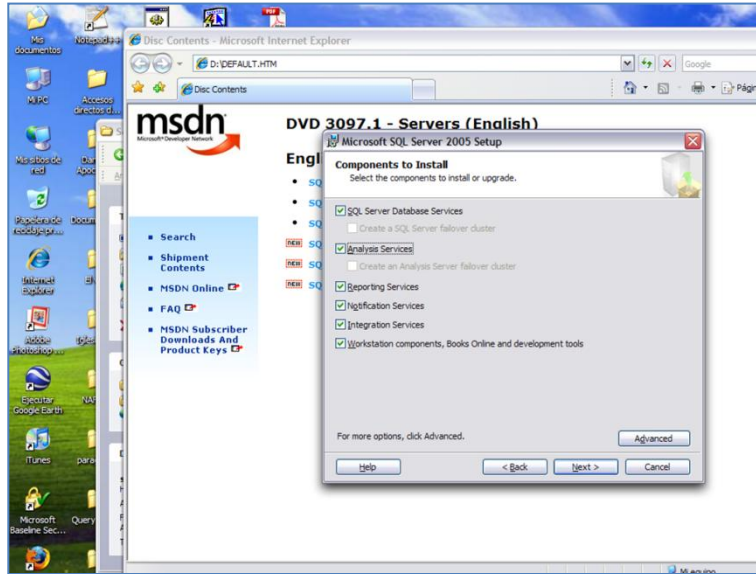


Figura 4 - Continuación instalación de SQL Server

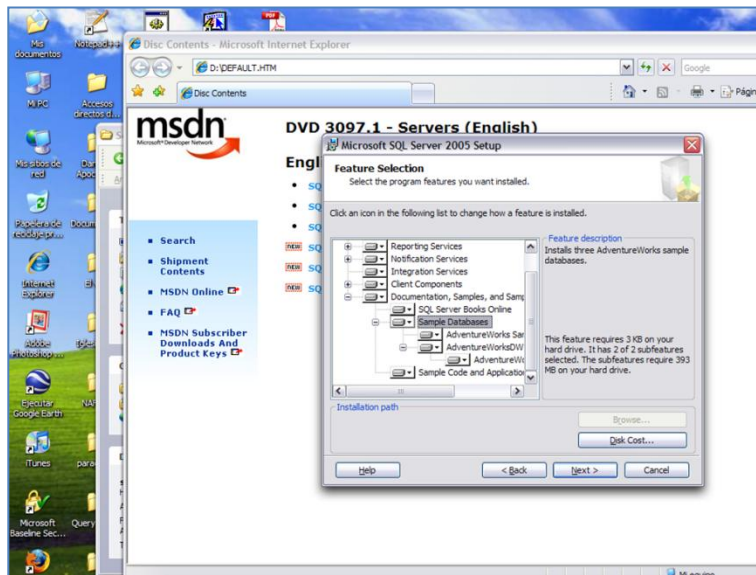


Figura 5 - Continuación de instalación ...

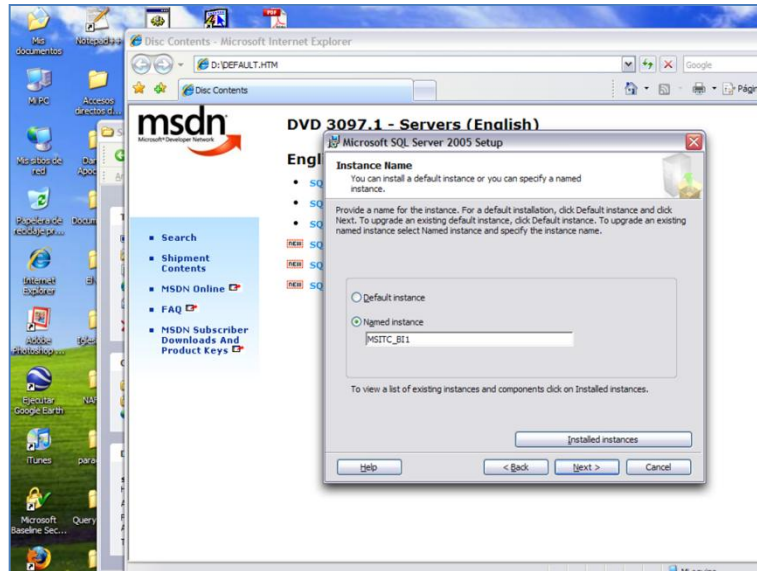


Figura 6 - Repaso de conceptos BI

Repaso de Conceptos BI

- Datos, Información y Conocimiento
- OLAP – OLTP
- Datawarehouse y Datamart
- ETL
- MOLAP, ROLAP, HOLAP
- Datamining
- Por qué BI?

Figura 7 - Datos e información

Datos, Información, Conocimiento

- Datos: elementos mínimos de información que por sí mismos son irrelevantes para la toma de decisiones. Ejemplo: 23,423, Juan Pérez, La Estrella.
- Información: datos procesados y que tienen significado, ejemplo: Ventas del mes: 23,423 unidades, Vendedor: Juan Perez, Marca: La Estrella.
- Información = Datos + Contexto (añadir valor) + Utilidad (disminuir la incertidumbre)

Figura 8 - Conocimiento

Datos...

- Conocimiento: El conocimiento es una mezcla de experiencia, valores, información y *know-how* que sirve como marco para la incorporación de nuevas experiencias e información, y es útil para la acción, ejemplo: Ventas del mes 23,423, 15% más alto que el mes anterior, Juan Pérez atiende a clientes mayoristas, La Estrella marca a la que se le incrementó 50% de fondos en publicidad, respecto del mes anterior.

Figura 9 - Pirámide de conocimiento

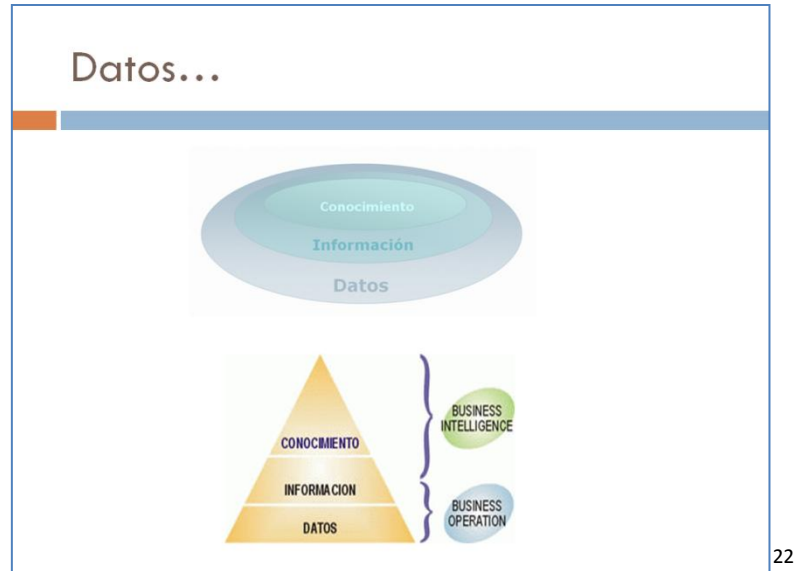


Figura 10 - OLTP

OLTP

- OLTP – *OnLine Transaction Processing*
- El acceso a los datos está optimizado para tareas frecuentes de lectura y escritura.
- Los datos se estructuran según el nivel aplicación.
- Los formatos de los datos no son necesariamente uniformes en los diferentes departamentos .
- El historial de datos suele limitarse a los datos actuales o recientes.

²² http://www.sinnexus.com/business_intelligence/index.aspx consultada en agosto 2008

Figura 11 - OLAP

OLAP – OnLine Analytical Processing

- El acceso a los datos suele ser de sólo lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o eliminaciones.
- Los datos se estructuran según las áreas de negocio, y los formatos de los datos están integrados de manera uniforme en toda la organización.
- El historial de datos es a largo plazo, normalmente de dos a cinco años.
- Las bases de datos OLAP se suelen alimentar de información procedente de los sistemas operacionales existentes, mediante un proceso de extracción, transformación y carga (ETL).

Figura 12 - Datawarehouse y datamart

Datawarehouse y Datamart

- *Datawarehouse*, base de datos corporativa, que integra y depura información de una o más fuentes distintas, permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta.
- *Datamart*, base de datos departamental, especializada en almacenar información de un área de negocios específica, generalmente es una sección de un *datawarehouse*.

Figura 13 - Características de *datawarehouse*

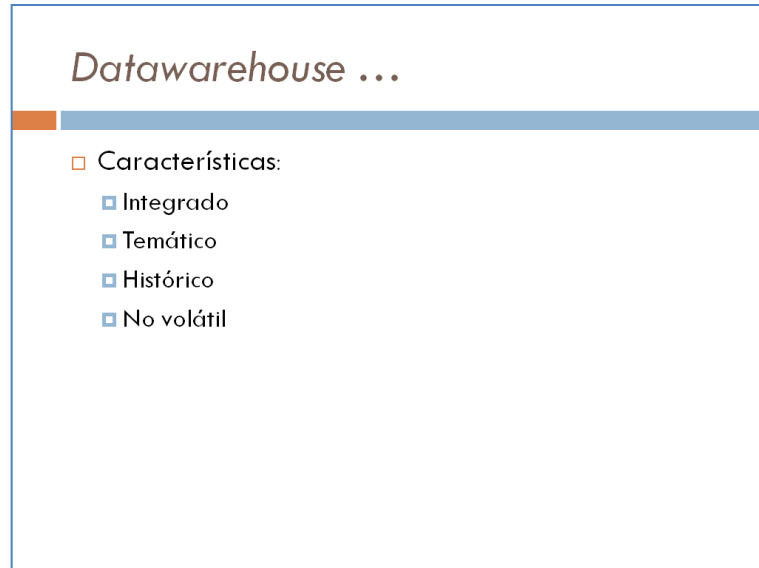


Figura 14 - Portada sesión 2, curso 1



Figura 15 - Agenda sesión 2, curso 1

Agenda

- Repaso de Conceptos de BI
- Visión general de un cubo

Figura 16 - ETL

ETL

- **Extracción:** obtención de información de las distintas fuentes tanto internas como externas.
- **Transformación:** filtrado, limpieza, depuración, homogeneización y agrupación de la información.
- **Carga:** organización y actualización de los datos y los metadatos en la base de datos.

Figura - 17 Proceso de ETL

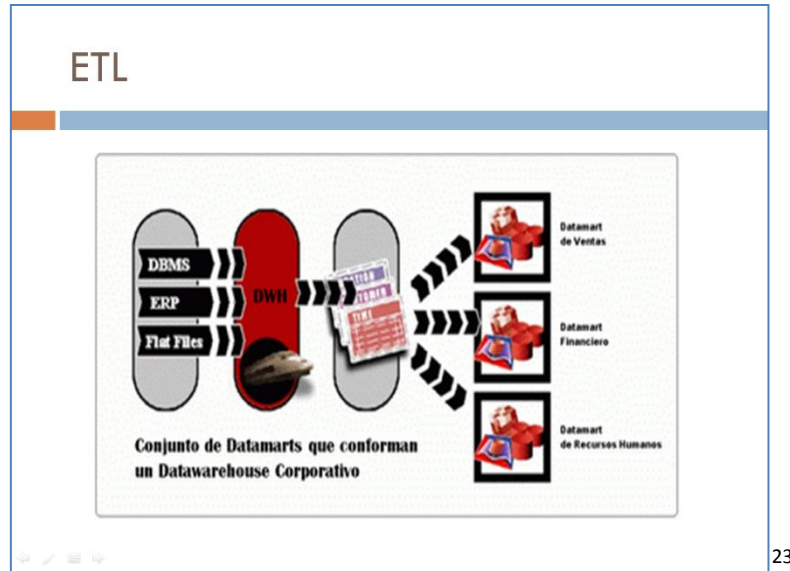


Figura 18 - MOLAP, ROLAP y HOLAP

MOLAP, ROLAP, HOLAP

- MOLAP – OLAP Multidimensional, calcula agregaciones y estructuras en motores multidimensionales, tiene las combinaciones pre-calculadas, requiere mucho espacio.
- ROLAP – OLAP Relacional, se basa en un motor de base de datos relacional, y realiza los cálculos de las agregaciones en el momento que se solicitan.
- HOLAP – OLAP Híbrido, una combinación entre los dos esquemas anteriores

²³ http://www.sinnexus.com/business_intelligence/index.aspx consultada en agosto 2008

Figura 19 - Datamining

Datamining

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

Figura 20 - Proceso de datamining


Datamining

- Determinación de los objetivos.
- Pre-procesamiento de los datos.
- Determinación del modelo
- Análisis de resultados

Figura 21 - ¿Por qué BI?

Por qué BI?

- **Observar** ¿qué está ocurriendo?
- **Comprender** ¿por qué ocurre?
- **Predecir** ¿qué ocurriría?
- **Colaborar** ¿qué debería hacer el equipo?
- **Decidir** ¿qué camino se debe seguir



24

Figura 22 - Continuación ¿Por qué BI?

Por qué BI?

- Solución Tecnológica
 - Centralizar, depurar y afianzar datos
 - Descubrir información no evidente para las aplicaciones actuales
 - Optimizar el rendimiento de los sistemas
- Ventaja Competitiva
 - Seguimiento real del plan estratégico
 - Aprender de errores pasados
 - Mejorar la competitividad
 - Obtener el verdadero valor de las aplicaciones de gestión

²⁴ http://www.sinnexus.com/business_intelligence/index.aspx consultada en agosto 2008

Figura 23 - Bibliografía sesiones 1 y 2

A presentation slide titled "Bibliografía" with a blue header bar and an orange bar on the left. It contains a list of five URLs. At the bottom left, there are small navigation icons: a left arrow, a pencil, a list icon, and a right arrow.

Bibliografía

- <http://es.wikipedia.org/wiki/ETL>
- <http://es.wikipedia.org/wiki/Datamart>
- <http://es.wikipedia.org/wiki/OLAP>
- <http://es.wikipedia.org/wiki/OLTP>
- http://www.sinnexus.com/business_intelligence/index.aspx

Figura 24 - Portada sesión 3, curso 1

A presentation slide with a dark brown background. The text "LABORATORIO INTELIGENCIA DE NEGOCIOS 1" is centered in white. At the bottom, there is a blue bar with the text "Sesión 3, 30-Ago-2008" and an orange bar on the left.

**LABORATORIO INTELIGENCIA
DE NEGOCIOS 1**

Sesión 3, 30-Ago-2008

Figura 25 - Agenda sesión 3, curso 1

Agenda

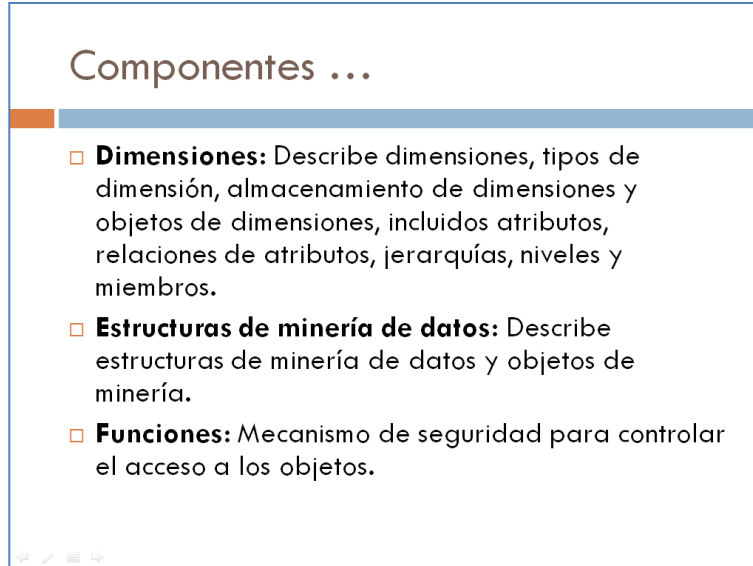
- Componentes de Analysis Services
- Usos típicos de Integration Services
- Ejemplos de tipos de ETL

Figura 26 - Componentes de SSAS

Componentes de SSAS

- **Orígenes de Datos:** Establece los mecanismos de conexión hacia el origen relacional de los datos.
- **Vistas de Origen de Datos:** Describe una vista de origen de datos, un esquema relacional y consultas asociadas utilizadas para modelar uno o más orígenes de datos, en Analysis Services.
- **Cubos:** Describe cubos y objetos de cubo, lo que incluye medidas, grupos de medida, relaciones de uso de dimensiones, cálculos, indicadores clave de rendimiento, acciones, traducciones, particiones y perspectivas.

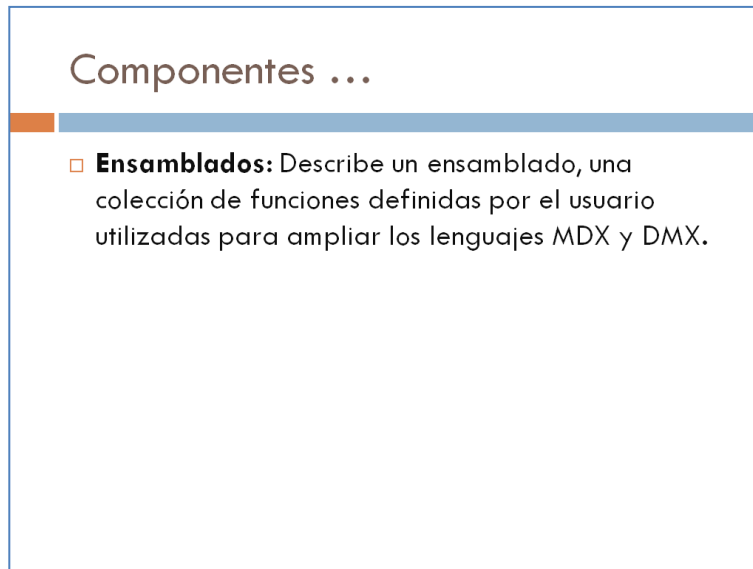
Figura 27 - Continuación de componentes de SSAS



Componentes ...

- **Dimensiones:** Describe dimensiones, tipos de dimensión, almacenamiento de dimensiones y objetos de dimensiones, incluidos atributos, relaciones de atributos, jerarquías, niveles y miembros.
- **Estructuras de minería de datos:** Describe estructuras de minería de datos y objetos de minería.
- **Funciones:** Mecanismo de seguridad para controlar el acceso a los objetos.

Figura 28 - Continuación de componentes de SSAS



Componentes ...

- **Ensamblados:** Describe un ensamblado, una colección de funciones definidas por el usuario utilizadas para ampliar los lenguajes MDX y DMX.

Figura 29 - Usos típicos de SSIS

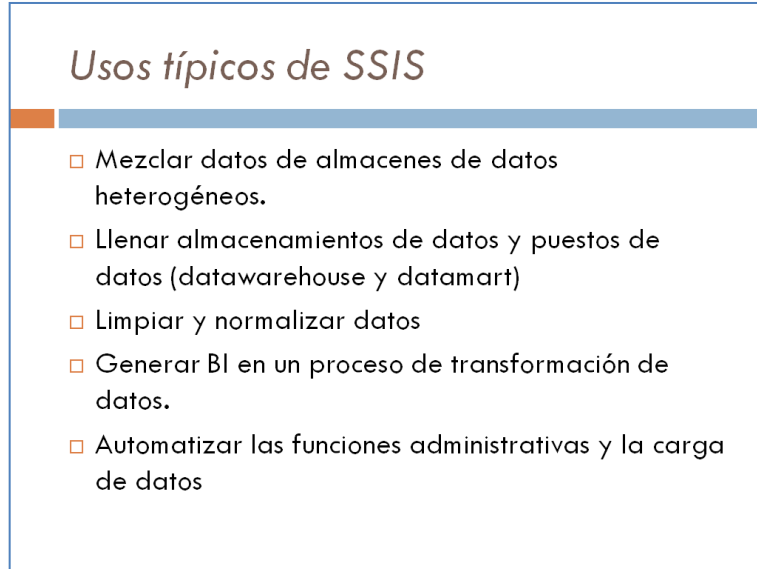


Figura 30 - Ejemplos de tipos de ETL

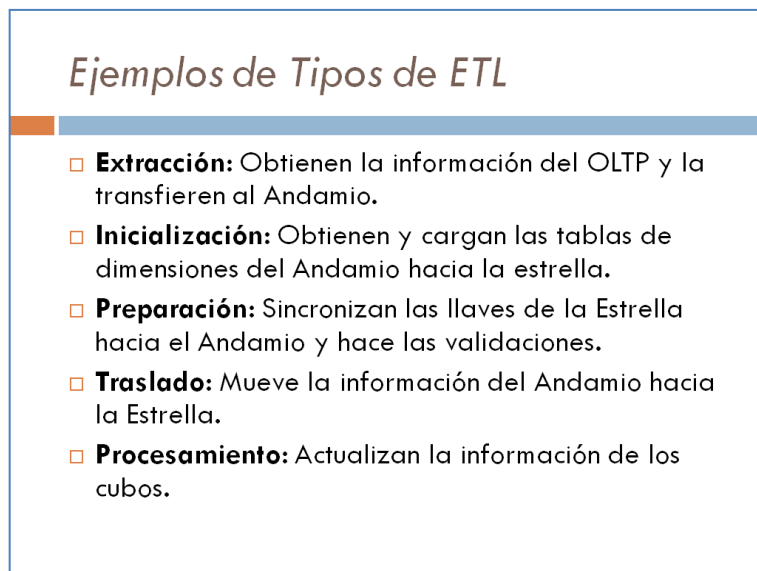


Figura 31 - ETL de extracción

ETL - Extracción

- En Andamio existe forma de relacionar el registro con el registro equivalente en el OLTP.
- Dimensiones: Revisa existencia de registro en Andamio, si es nuevo lo transfiere, si es distinto actualiza y marca el registro, si está igual, lo deja como tal.
- Hechos: Obtiene datos en rango de tiempo según la periodicidad definida, convirtiendo los id's del OLTP de una vez a los id's del Andamio.

Figura 32 - ETL de inicialización

ETL - Inicialización

- Aplican a las tablas de dimensiones
- Realizan procesos de consolidación (“aplanan” tablas)
- Filtran los registros que han sido modificados.
- Actualizan la información, proveniente del OLTP en la Estrella
- Crean valores “Default” para posibles valores nulos.

Figura 33 - ETL de preparación

ETL - Preparación

- Se realizan sobre los hechos.
- Realizan pre-cálculos necesarios para la estrella, (conversiones)
- Obtienen el id de la estrella, para cada uno de los campos asociados, en base al id del andamio.
- Verifican la ausencia de nulos
- Se realizan validaciones iniciales, cantidad de registros, suma de montos, razonabilidad de las conversiones, etc.

Figura 34 - ETL de carga

ETL - Carga

- Una vez que los datos en los hechos están preparados y validados.
- Trasladan la información desde el Andamio hacia la Estrella.
- Eliminan la información trasladada del Andamio, para estar listos para la próxima carga.
- Si se permite volver a cargar ciertos datos, deben ser capaces de actualizar esa información en la Estrella (más óptimo: borrar y volver a trasladar)

Figura 35 - ETL de procesamiento

ETL - Procesamiento

- Son los encargados de refrescar la información, ya actualizada en la Estrella (DW) hacia los cubos.
- Refrescan las dimensiones, si hubo cambios drásticos, re-construirlas, sino actualización incremental.
- Procesar cubos y modelos de minería de datos, puede ser incremental, refrescamiento o re-construcción.

Figura 36 - Portada sesión 1 y 2, del curso 2

**LABORATORIO INTELIGENCIA
DE NEGOCIOS 2**

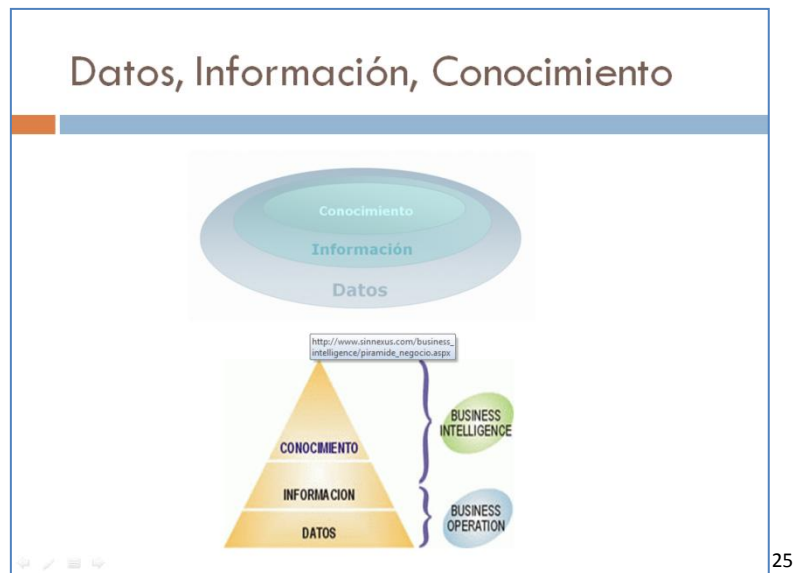
Sesión 1, 10-ene-09, Sesión 2, 17-ene-09

Figura 37 - Agenda sesiones 1 y 2, curso 2

Agenda

- Repaso Datos, Información, Conocimiento
- Descubriendo el conocimiento (KDD)
- Técnicas de KDD
- Minería de Datos

Figura 38 - Repaso datos, información y conocimiento



²⁵ http://www.sinnexus.com/business_intelligence/index.aspx consultada en agosto 2008

Figura 39 - Descubrir el conocimiento

Descubrir el Conocimiento (KDD)

- Se define como “la extracción no trivial de información implícita, desconocida y potencialmente útil de los datos”.
- El proceso de KDD toma los resultados tal como vienen los datos, los transforma en información útil y entendible.
- Procesando grandes cantidades de datos crudos, identifican los patrones significativos y relevantes y los presentan como conocimiento apropiado para satisfacer las metas del usuario

Figura 40 - Técnicas de KDD

Técnicas de KDD

- Método de Clasificación
 - Es el más usado, agrupa los datos de acuerdo a similitudes o clases.
- Método Probabilístico
 - Utilizando modelos de representación gráfica, se basa en las probabilidades e independencia de los datos, puede usarse en sistemas de diagnóstico, planeación y sistemas de control

Figura 41 - Continuación técnicas de KDD

Técnicas

- Método Estadístico
 - ▣ Usa la regla del descubrimiento y se basa en las relaciones de los datos, usado para generalizar los modelos en los datos y construir las reglas de los modelos nombrados.
- Método Bayesiano
 - ▣ Es un modelo gráfico, usado frecuentemente las redes de Bayesian cuando la incertidumbre se asocia con un resultado que puede expresarse en términos de probabilidad, usado en sistemas de diagnóstico.

Figura 42 - Minería de datos

Minería de Datos

- Es un mecanismo de explotación, consistente en la búsqueda de información valiosa en grandes volúmenes de datos.
- Otra definición: es el análisis de archivos y bitácoras de transacciones, trabaja a nivel del conocimiento con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones.
- La MD está muy ligada a los Data Warehouse

Figura 43 - División de la minería de datos

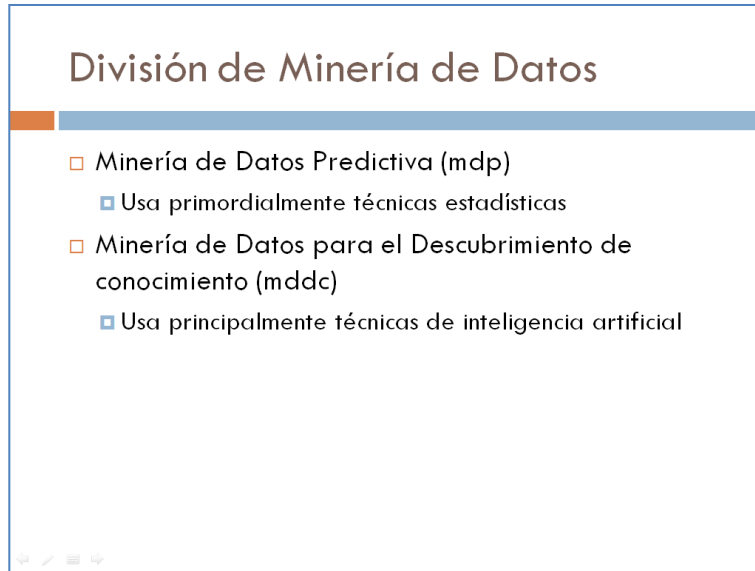


Figura 44 - Aplicaciones de minería de datos



Figura 45 - Técnicas de minería de datos

Técnicas de Minería de Datos

- Análisis preliminar de datos usando herramientas de consulta.
- Técnicas de visualización
- Redes neuronales artificiales
- Reglas de asociación
- Algoritmos genéticos
- Redes Bayesianos
- Árboles de decisión

Figura 46 - Algoritmos de minería de datos

Algoritmos de Minería de Datos

- Supervisados o predictivos
 - ▣ Predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos.
- No supervisados o del descubrimiento del conocimiento
 - ▣ Con estos algoritmos se descubren patrones y tendencias en los datos actuales. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio de ellas.

Figura 47 - Proceso de minería de datos

Proceso de la Minería de Datos

- **Determinación de los objetivos**
 - ▣ Delimitar los objetivos que el cliente desea.
- **Pre procesamiento de los datos**
 - ▣ Se refiere a la selección, limpieza, enriquecimiento, reducción y transformación de las bases de datos
- **Determinación del modelo**
 - ▣ Análisis estadístico de los datos y visualización gráfica de los mismos como una primera aproximación
- **Análisis de los resultados**
 - ▣ Verificar la coherencia de los resultados obtenidos y compararlos con los resultados estadísticos y gráficos

Figura 48 - Extensiones de minería de datos

Extensiones de Minería de Datos

- **Web Mining**
 - ▣ Consiste en aplicar técnicas de minería de datos a documentos y servicios de la web, procesando los logs para producir información significativa.
- **Text Mining**
 - ▣ Se refiere a examinar una colección de documentos y descubrir información no contenida en ningún documento individual de la colección.
 - ▣ Dado que el 80% de la información de una empresa se almacena en forma de documentos, existen técnicas que apoyan el TM

Figura 49 - Por qué usar minería de datos

Por qué usar Minería de Datos?

- Contribuye a la toma de decisiones tácticas y estratégicas.
- Proporciona poder de decisión a los usuarios del negocio, y es capaz de medir las acciones y resultados de la mejor forma.
- Genera modelos descriptivos: permite a empresas, explorar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales.
- Genera modelos predictivos: permite que relaciones no descubiertas través del proceso del DM sean expresadas como reglas de negocio.

Figura 50 - Portada sesión 3, curso 2

LABORATORIO INTELIGENCIA DE NEGOCIOS 2

Sesión 3, 30-ene-09

Figura 51 - Agenda sesión 3, curso 2

Agenda

- Proceso de Datamining
- Datamining en BDD
- Oracle Data Mining(ODM)
- Funciones de ODM
- Funciones soportadas por ODM

Figura 52 - Proceso de *datamining*

Proceso de Data Mining

- En la vida real, los datos por lo general se encuentran “sucios”, con datos incorrectos o ausentes.
- Se deben “limpiar” antes de utilizarlos, filtrando, normalizando, tomando muestras, transformando en varias direcciones.
- Cerca del 80% del esfuerzo en un proyecto de DM es invertido en la preparación de los datos.
- Estos datos son la entrada para algún algoritmo de DM

Figura 53 - Continuación proceso de *datamining*

Proceso de Data Mining...

- Los modelos son construidos, probados, validados, administrados e implementados en un ambiente apropiado.
- Los resultados del DM pueden ser post-procesados como parte de cálculos específicos y por lo tanto almacenados en una base de datos permanente.
- Puede involucrar la transferencia entre servidores, repositorios de datos, aplicaciones y herramientas, conversiones de formatos, etc.

Figura 54 - *Datamining* en BDD

Data Mining en BDD

- Eliminando o reduciendo estos obstáculos, se puede ejecutar el DM con mayor frecuencia.
- Se pueden utilizar datos más actualizados
- Se reduce el movimiento de datos, con lo que se reduce el tiempo total del procesamiento del DM.
- Si los datos no abandonan la base de datos, se mantiene la seguridad sobre los datos.

Figura 55 - Oracle (R) *Datamining*

Oracle Data Mining

- ODM Integra la minería de datos, dentro de la base de datos Oracle.
- Los algoritmos de DM operan nativamente sobre tablas relacionales o vistas, lo que elimina ETL para una herramienta especializada.
- Las tareas de DM pueden ejecutarse asíncronamente e independientes de alguna interface como parte de una bdd normal.
- Las herramientas de ODM pueden ejecutarse en línea con interfaces en Java, o bien con PL/SQL

Figura 56 - Funciones de ODM

Funciones de ODM

- Supervisadas (Directas)
 - ▣ Usadas para predecir un valor
 - ▣ Requieren la especificación de un objetivo, el cual puede tener atributos binarios (compra, no compra) o bien una lista de alternativas (color de sweater, rangos de salarios, etc)
 - ▣ Naive Bayes para clasificaciones es un algoritmo para minería supervisada

Figura 57 - Continuación de funciones de ODM

Funciones de ODM

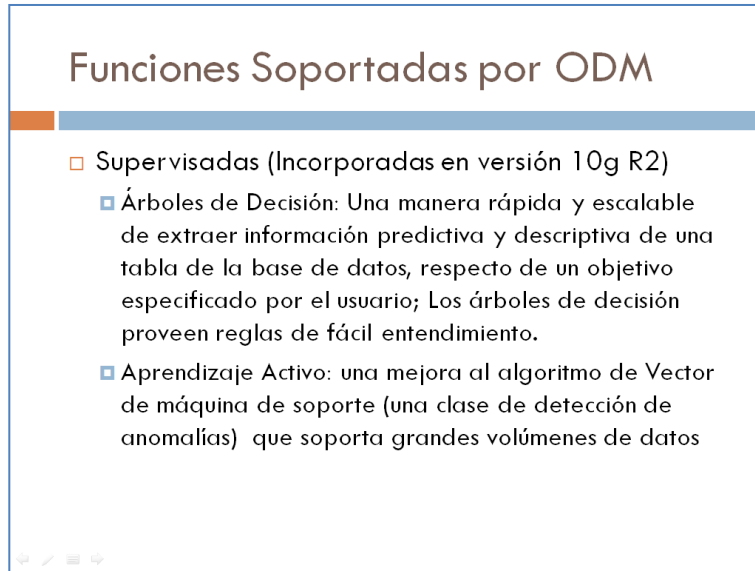
- No Supervisadas (Indirectas)
 - ▣ Usadas para encontrar estructuras intrínsecas, relaciones o afinidades en los datos.
 - ▣ No usan un objetivo.
 - ▣ Los algoritmos de "Clustering", pueden ser usados para encontrar las agrupaciones naturales en los datos.
- También se pueden clasificar como Predictivos o Descriptivos
 - ▣ Predictivos: Clasificación y Regresión
 - ▣ Descriptivos: Conjunto de características de los datos

Figura 58 - Funciones soportadas por ODM

Funciones Soportadas por ODM

- Supervisadas
 - ▣ Clasificación: Agrupa los ítems en clases discretas y predice a qué clase pertenece un ítem.
 - ▣ Regresión: La aproximación y la previsión de valores continuos
 - ▣ Importancia del atributo: Identificar los atributos que son más importantes en la predicción de resultados
 - ▣ Detección de anomalías: Identificar los elementos que no cumplan las características de los datos "normales" (outliers)

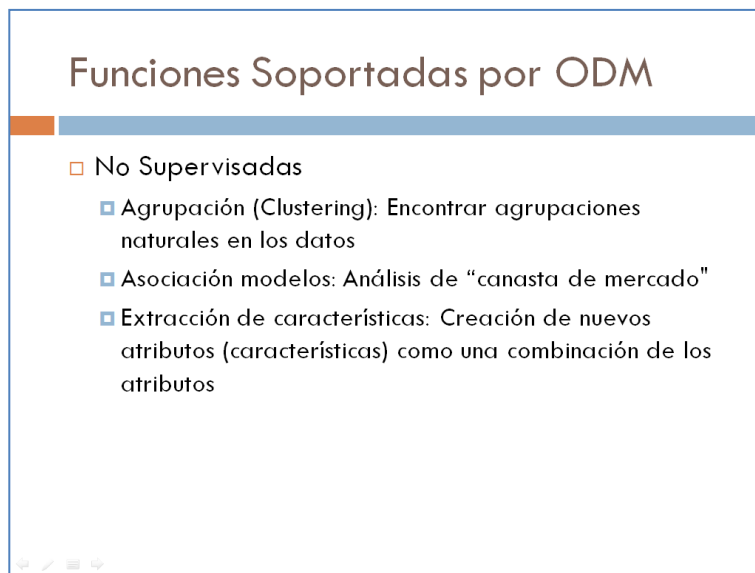
Figura 59 – Continuación de funciones soportadas por ODM



Funciones Soportadas por ODM

- Supervisadas (Incorporadas en versión 10g R2)
 - ▣ Árboles de Decisión: Una manera rápida y escalable de extraer información predictiva y descriptiva de una tabla de la base de datos, respecto de un objetivo especificado por el usuario; Los árboles de decisión proveen reglas de fácil entendimiento.
 - ▣ Aprendizaje Activo: una mejora al algoritmo de Vector de máquina de soporte (una clase de detección de anomalías) que soporta grandes volúmenes de datos

Figura 60 - Continuación de funciones soportadas por ODM



Funciones Soportadas por ODM

- No Supervisadas
 - ▣ Agrupación (Clustering): Encontrar agrupaciones naturales en los datos
 - ▣ Asociación modelos: Análisis de “canasta de mercado”
 - ▣ Extracción de características: Creación de nuevos atributos (características) como una combinación de los atributos

Figura 61 - Portada sesión 4, curso 2



Figura 62 - Agenda sesión 4, curso 2

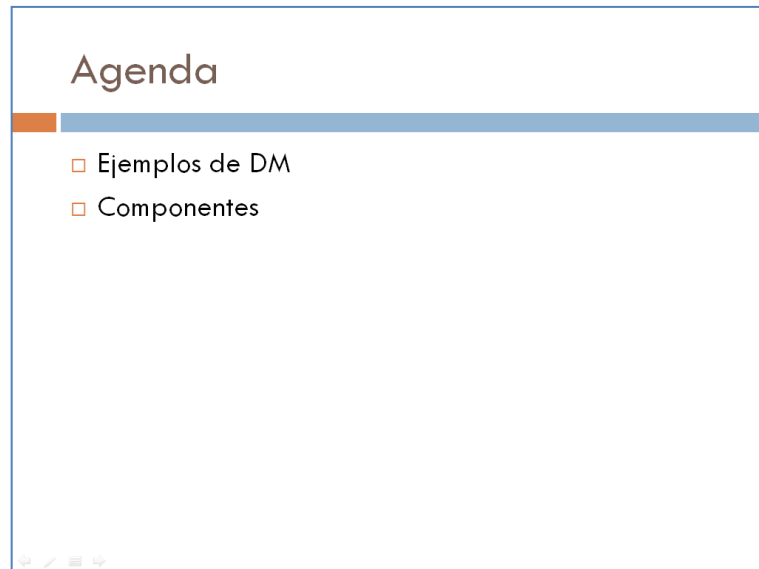


Figura 63 - Ejemplo 1 de ODM

Ejemplos de DM

- Problema: Un vendedor al detalle desea incrementar sus ingresos, identificando sus principales clientes potenciales, para crear incentivos para ellos. También desea una guía en su almacenamiento, determinando los productos que más frecuentemente se compran juntos.
- Solución: Un modelo de Clasificación se puede construir para determinar los clientes que están “dispuestos” más de un 75% de gastar más de \$1000 el próximo año. Un modelo de Reglas de Asociación para crear un análisis de canasta.

Figura 64 - Ejemplo 2 de ODM

Ejemplos...

- Problema: Una agencia gubernamental desea métodos más rápidos y confiables para identificación de posibles actividades fraudulentas para futuras investigaciones.
- Solución: Crear modelos de Clasificación, Clustering y modelos de detección de Anormalidades para marcar los casos “sospechosos”

Figura 65 - Ejemplo 3 de ODM

Ejemplos...

- Problema: Un investigador bioquímico desea trabajar con miles de atributos asociados con una investigación de la efectividad de una droga.
- Solución: Una función de Importancia de atributos para reducir el número de factores a un subconjunto manejable de atributos.

Figura 66 - Ejemplo 4 de ODM

Ejemplos

- Problema: Una compañía hipotecaria desea incrementar los ingresos, reduciendo el tiempo requerido para la aprobación de los préstamos.
- Solución: Un modelo de Regresión puede predecir el mejor valor para una casa, eliminando la necesidad de una inspección en sitio.

Figura 67 - Aplicación de ejemplos de ODM

Ejemplos...

- Si tienen un caso similar a estos, Oracle Data Mining puede ayudarles a encontrar la solución.
- Si se intenta resolver estos problemas con ODM, se puede estar seguro que el conocimiento del negocio y su conocimiento de los datos de las áreas de negocios, son los factores más importantes del proceso.
- ODM automatiza la mecánica de construcción, pruebas y aplicación de modelos, de manera que nos ocupemos de los aspectos de negocios del problema y no de detalles estadísticos y matemáticos

Figura 68 - Componentes de ODM

Componentes

- ODM es accesible desde tres distintas interfaces, cada una diseñada para un tipo de usuario distinto.
- Oracle Data Mining Predictive Analytics (PA), es un paquete que contiene dos programas –Predecir y Explicar- cada uno requiriendo solamente la data de entrada