**Electiva III: Business Intelligence & Big Data**

Docentes:

* Marcela Andrea Vera
* Lorena D’iorio
* Alejandro Aguirre

Horario: 17:10 a 18:30 (Teoría), 18:30 a 20:45 (Práctica).

Unidades:

1. Bussines Intelligence y Sistemas de Soporte de Decisión
2. Datawarehouse – Herramientas de Implementación.
3. Big Data – Características y técnicas.

Bibliografía Básica:

* Building the Data Warehouse – Inmon, W.
* Data Mining for the Masses – North, M.
* The Data Warehouse Lifecycle Toolkit – Kimball, R.
* Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data – Zikopolous Paul, Deroos Dirk.
* Pattern-Based Strategy: Getting Value from Big Data – Yvonne Genovese.

Requisitos para regularizar

* 80% de asistencia.
* Un parcial o su recuperatorio con nota mayor a 40.
* Entrega y aprobación de todos los TPs.
* Participación activa en los fotos y actividades del aula virtual.

Requisitos para promocionar:

* Nota mayor o igual a 80 (se puede recuperar).

Requisitos para aprobar:

* Alumnos regulares: examen final teórico-práctico escrito.
* Alumnos promocionados: se mantiene la promoción hasta el inicio del dictado de clases del primer cuatrimestre del año siguiente al cursado.

**Las Organizaciones Industriales – Una primera aproximación**

Introducción

* 1820-1840: durante más de 100 años las empresas han producido bienes y servicios en un entorno de crecimiento y desarrollo.
* Fueron inventados y fabricados los autos, los ferrocarriles y el teléfono, entre muchos otros.
* Los principios sobre los cuales se organizaban, se adaptaban muy bien a las condiciones de la época.
* Las Organizaciones actualmente afrontan nuevas complicaciones que complican su eficiencia.
* Las raíces organizacionales de las empresas de hoy responden al prototipo de fábrica de Adam Smith.

1776 – Adam Smith

* Filósofo y Economista.
* Precursor de la consultoría de negocios.
* Analiza las industrias a la luz de las nuevas posibilidades ofrecidas por la Revolución Industrial.
* Inicia sus estudios en una fábrica de alfileres.
* Escribe “Las riquezas de las naciones” donde detalla su “principio de la división del trabajo”. “Un hombre estira el alambre, otro lo endereza, el tercero lo corta, el cuarto le saca punta, el quinto lo pule por encima para recibir la cabeza, etc…”
* Ventajas para las industrias:
  + Surgen trabajadores especializados que realizaban un solo paso de la fabricación.
  + Mayor velocidad en la realización del trabajo. Se ahorra el tiempo de pasar de una tarea a la otra.
  + Facilitó la invención de máquinas que realizaban el trabajo de muchas personas.
  + Disminuyó los requerimientos de capacitación de los empleados.
  + Facilitó la incorporación de nuevos trabajadores.
  + Promovía el mejoramiento de las prácticas de los trabajadores en tareas específicas.
  + Aumentaba la producción.
* Problemas para los trabajadores:
  + Cada trabajador solo realiza una parte del trabajo.
  + No tenía la visión general de la tarea.
* Entorno
  + En 1776 no había mercado para aprovechar estas ventajas.
  + No había buenos caminos ni medios de comunicación.
  + Solo se podía acceder a mercados locales regionales.

1820 – El Ferrocarril

* Se inventan los ferrocarriles.
* Se mejoran los medios de comunicación.
* Se impulsa el desarrollo económico. Las empresas podían llegar a mercados más lejanos.
* Gran crecimiento demográfico.
* Consecuencias:
  + Surge la Tecnología de la Administración de Negocios. Tecnología Blanda.
  + Fue desarrollada para evitar los choques de trenes en una sola vía férrea por la que circulaban en ambos sentidos.
  + Se desarrollan los cimientos de la Burocracia Moderna.
  + Se establecieron líneas de autoridad y dependencia.
  + Los operarios trabajaban en función de reglas preestablecidas, programadas.
  + Se desarrollan planes de contingencia.
  + Mediante la tecnología de la administración, las empresas podían crecer más allá del control de un solo individuo.

Principios del Siglo XX – Henry Ford

* Henry Ford refinó la división del trabajo de Adam Smith.
* Cada trabajador instalaba una sola pieza de una forma preestablecida.
* La mayor innovación fue llevar el trabajo hasta el trabajador, y no a la inversa.
* Problemas:
  + La división extrema comenzaba a complicar la coordinación del trabajo.
  + No era fácil combinar los pedazos de resultados para obtener el producto o servicio final.
  + El sistema administrativo no soportaba la magnitud que habían logrado las industrias.
  + Ni Ford, ni William Durant, el fundador de General Motors, habían podido manejar las enormes y fragmentadas organizaciones.

Principios del Siglo XX – Alfred Sloan

* Sucedió a William Durant como presidente de GM.
* Impulsó el cambio anual de estilos de los automóviles. Derivó el concepto de obsolescencia programada.
* En Ford, la administración había sido más simple porque había menos modelos.
* En GM había más variedad de autos y modelos
  + La compañía producía demasiados coches de un modelo.
  + Tenían que suspender la producción porque no tenían materia prima.
* Cuando asume, Sloan crea el sistema que hoy se conoce como “Producción en Serie”.
* Aplicó el principio de Adam Smith a la administración.
* Genera una administración para esa “producción en serie”.
* Alcanzó notoriedad por gestionar diversas operaciones mediante el control de estadísticas y parámetros financieros.
* La Administración de Sloan
  + Creó divisiones más pequeñas descentralizadas y controlables. Una división para cada modelo de automóvil.
  + Los ejecutivos no requerían conocimientos específicos de ingeniería o manufactura.
  + Necesitaban pericia financiera, debían estudiar los números, ventas, ganancias, niveles de inventario, etc.

1960 – La Planificación

* Robert McNamara en Ford, Harold Geneen en ITT, Regionald Jones en General Electric.
* Desarrollan las técnicas de planificación.
  + Determinación de los negocios a los que quería dedicarse la empresa.
  + Determinación del capital que debía destinarse.
  + Determinación de utilidades esperadas.
* Consecuencia:
  + Surgen controladores (sinónimo de “indicador”), auditores y planificadores para llevar adelante estos planes.
  + Recaban información de inventario y actividades para ajustar la planificación.
  + Se intensifican y perfeccionan las técnicas para presupuestar, controlar y planificar.

Finales de la II Guerra Mundial – Situación

* Demanda insaciable de bienes y servicios.
* La estructura piramidal era muy útil por su escalabilidad.
* Cuando una compañía quería crecer, agregaba trabajadores en la base del organigrama.
* La principal preocupación de los ejecutivos desde el punto de vista operativo era la capacidad.

Nuestros días

* Los procesos se fueron complicando.
* La cantidad de tareas crecía.
* Aumentó el personal en los niveles intermedios del organigrama.
* Mayor distancia entre la administración y los clientes.
* Avance de la tecnología.
* Desaparición de las fronteras.

**Nuevos Tiempos**

Hay 3 fuerzas básicas que impulsan a las compañías a producir un cambio.

Tres “C”:

* Clientes
* Competencia
* Cambio

Los clientes asumen el mando

* Cambia la fuerza dominante en la relación vendedor-cliente.
* Antes: los clientes se adaptaban. La oferta era poca.
* Hoy: la empresa se adapta al cliente. La oferta es mucha.
* Los clientes exigen personalización.
* Mayor diversificación.
* La tecnología es la gran impulsora del cambio.

La competencia se intensifica

* Se diversifica sobre bases competitivas totalmente distintas.
* No existen territorios protegidos de competencia extranjera.
* Un competidor eficiente sube el umbral competitivo para los demás.
* Nuevos competidores.
* La tecnología cambia la naturaleza de la competencia.

El cambio se vuelve constante

* General y permanente.
* La rapidez del cambio tecnológico promueve la innovación.
* Cortos ciclos de vida de los productos.
* Disminuye el tiempo disponible para innovar.

Conclusión

* La división del trabajo y la especialización de las tareas estratificaron los procesos separando aquellas actividades relacionadas a la producción de aquellas asociadas a la gestión.
* Podemos categorizar entonces a los procesos en algún punto en un continuo entre dos extremos:
  + Los completamente estructurados.
  + Los completamente no estructurados.
* Paradójicamente, mientras la evolución de las organizaciones se mueve hacia ambientes menos estructurados, los sistemas han evolucionado en general dentro de los ambientes estructurados.
* El volumen de información empresarial se encuentra en un crecimiento constante. Muchos datos están dejando de ser estudiados debido a la incapacidad de los sistemas actuales para procesarlos.
* Obsesión con la actividad más que con el resultado.
* Si una organización está en un nicho exitoso, seduce la competencia. Hoy la competencia tiene barreras de entrada más débiles que años atrás.
* La competencia baja el rendimiento del nicho.
* El éxito de las organizaciones depende de la innovación continua. Se trata de administrar sus recursos para ofrecer cosas nuevas y mantenerse exitoso lo más posible dentro del nicho, aprovechando el éxito para generar nuevas oportunidades.
* Se han generado y se generan muchos datos. El desafío es encontrar la información oculta y darle un significado.
* La tecnología hoy dejó de ser una facilitadora para ser una posibilitadora de la administración de la información de los procesos.
* La tecnología de la información está transformando la forma de hacer negocios y los negocios mismos.
* Las grandes innovaciones de la humanidad fueron tecnologías genéricas. Han producido éxito en la medida que fueron aplicadas exitosamente. Requirieron de siglos para ofrecer su máxima utilidad.
* Hoy indudablemente la innovación pasa por la información en todas sus formas.
* Juegan entonces un papel preponderante los desarrollos que se hagan para aplicar esta innovación.
* Seguramente no va a tomar siglos obtener los máximos beneficios de estas tecnologías, pero todavía no sabemos el impacto final que va a tener.
* Se trata de estar atento y al día.

¿Qué es Business Intelligence (BI)?

Conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización.

Business Intelligence (BI)

* BI busca obtener conocimiento del negocio.
* Es capturada la información de las distintas áreas de la empresa con el objetivo de establecer estrategias fortalezas y debilidades.
* Marcan una evolución de los sistemas de información tradicionales (sistemas operacionales [SO]), pero no sustitutiva sino complementaria.

Dato: representación simbólica de un atributo. NO tiene un valor semántico. *Ejemplo:* sueldo correspondiente al mes de enero de 2010 de un docente universitario con cargo de adjunto por una dedicación semi-exclusiva.

Información: conjunto organizado de datos procesados. Tiene significado, relevancia y propósito. Los datos se convierten en información cuando su creador les añade significado. *Ejemplo:* es susceptible de adquirir un préstamo por un monto X a pagar en 12 cuotas mensuales.

Conocimiento: es una mezcla de experiencia, información y “saber hacer”. Marco para la incorporación de nuevas experiencias e información. Es útil para la acción. Se origina y aplica en la mente de los conocedores. *Ejemplo:* si tiene hijos, casa propia y supera los procesos administrativos de más de 2 meses son buenos candidatos para la entrega de préstamos para la adquisición de rodados. Se puede mejorar la tasa de referencia en un 35%.

Sistemas Operacionales vs Business Intelligence

*Sistemas Operacionales*

* Optimizados para operar.
* Normalizados para apoyar operaciones continuas de inserción, modificación y borrado de datos.

*Business Intelligence*

* Optimizadas para preguntar.
* Los datos están desnormalizados para apoyar consultas de alto rendimiento.

Características de un sistema de Business Intelligence

* Accesibilidad a la información. Los usuarios deben poder acceder con independencia de la procedencia.
* Apoyo en la toma de decisiones (gerencia). Ir más allá en la presentación de la información. Acceso a herramientas de análisis que permitan seleccionar y manipular sólo aquellos datos de interés.
* Orientación al usuario final. Se busca independencia entre los conocimientos técnicos y la capacidad para acceder a la información.

Procesos de decisión

* Implica elegir entre varias alternativas o cursos de acción con el propósito de alcanzar uno o varios objetivos definidos.
* No solo implica resolver problemas sino también investigar oportunidades de negocio.

Taxonomía de Simon (1977)

Clasifica a los procesos de decisión en un intervalo continuo entre dos extremos:

* Proceso de decisión estructurado -> Automatizable.
* Proceso de decisión NO estructurado -> NO automatizable.
* No estructurados: no tienen fases estructuradas. Se resuelven con la intuición humana.
* Semi-estructurados: tienen alguna fase estructurada. Se resuelven combinando procedimientos estándares y juicio humano.
* Estructurados: todas las fases son estructuradas. Se conocen los procedimientos para obtener la mejor solución.

**Soporte Informático para Procesos de Decisión Estructurados**

Aplicados a procesos:

* Repetitivos.
* Automatizables.
* Tipificables.

**Soporte Informático para Procesos de Decisión No Estructurados.**

Sistemas de soporte decisión (DSS)

* Son sistemas de información interactivos que ayudan al tomador de decisiones a utilizar datos y modelos para resolver problemas.
* No tiene capacidad para resolver problemas por sí solo.
* Tiene como propósito ayudar al decisor. No reemplazarlo.
* Ejemplos:
  + Soporte para Análisis Multidimensional (On Line Analitic Procesing [OLAP])
  + Soporte para procesos de Data Mining.
  + Soporte para pronósticos (Forecasting).
  + Soporte para Tablero Comando (Cuadro de Mando Integral - Balanced Score Card [BSC]).
  + Soporte para Gestión de Conocimiento.

**Conceptos Importantes**

**Data Warehouse**

¿Qué es?

* Es una base de datos (repositorio de información) corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta.

Características específicas

* Desnormalizada (está pensado para responder rápidamente consultas complejas).
* Optimizada para las consultas (está pensado para responder rápidamente consultas complejas).
* No volátil (es importante mantener toda la “historia”).

Características de su desarrollo

* Costoso (proceso incremental).
* Insume mucho tiempo.
* Identificación de fuentes (determinar, por ejemplo, en cuál de todas las fuentes X dato está más actualizado o es más fiel, sin errores).
* Se debe involucrar mucho con los procesos de toma de decisiones.

Posibles organizaciones que la utilizan/utilizarán

* Estatales.
* Privadas.

**Data Mining**

¿Qué es?

* DM se refiere a un conjunto de métodos estadísticos que proporcionan información (correlaciones o patrones) cuando se dispone de muchos datos (de aquí viene el nombre minería de datos). Datos + Estadísticas = Información.
* Proceso de analizar datos desde diferentes perspectivas con el objetivo de resumir los datos en segmentos de información útiles.
* El DM permite a los usuarios analizar los datos desde diferentes dimensiones o ángulos, categorizándolos y resumiendo relaciones identificadas.
* DM permite trabajar con grandes cantidades de observaciones (varios millones; información histórica de por lo menos, 5 años) sin ningún inconveniente.

Usos

* Comercio/Marketing:
  + Identificación de patrones de compra de los clientes.
  + Búsqueda de asociaciones entre clientes y características demográficas.
  + Promociones a partir de la identificación de comportamiento de diferentes estratos o categorizaciones de los clientes.
* Banca:
  + Detección de patrones de uso fraudulento de tarjetas de crédito.
  + Identificación de clientes leales.
  + Predicción de clientes con posibilidad de cambiar su afiliación.
* Medicina:
  + Identificación de terapias médicas adecuadas para diferentes enfermedades.
  + Estudio de factores de riesgo en diferentes patologías.

**Big Data**

¿Qué es?

* Almacenamiento de grandes cantidades de datos y a los procedimientos usados para encontrar patrones repetitivos dentro de esos datos (Terabytes a Petabytes).

¿Cuáles son las diferencias con un Data Warehouse?

* Aún más cantidades de datos.
* No estructurados y estructurados.
* Datos más complejos (texto, sonido).
* Se complementan.

Las tres “V” de Big Data:

* Velocidad: para procesar y a su vez hace referencia al hecho de que todo el tiempo se almacena información nueva.
* Volumen: grandes cantidades de datos con crecimiento exponencial.
* Variedad.

¿De dónde provienen los datos?

* Generados por las personas (sistemas operacionales).
* Por las transacciones de datos.
* E-marketing y web.
* Lectores biométricos y sensores.
* Machine to machine (M2M).

**Data Warehouse**

Base de datos para dar soporte a todos los procesos de decisión de la organización.

Características:

1. Orientación a temas
2. Integración
3. Relación con el tiempo
4. No volátil

1 - Orientación a Temas

* BD operacional: orientado a las aplicaciones, a los procesos/funciones operacionales. Ej.: Cuenta Corriente, Plazo Fijo, Tarjeta de Crédito.
* Data Warehouse: orientado a temas, a los procesos de decisión. Ej.: Ventas, Productos, Logística.
  + Es organizado en torno a temas principales
  + Excluye datos no usados para tomar decisiones
  + Los procesos de negocio no son considerados en el diseño del DW.

2 – Integración

Fuentes de Datos del DW

* Datos Internos: entorno operacional de la organización (Legacy Systems, DB operacionales, spreadsheets, etc.)
* Datos Externos

3 – Relación con el tiempo

* BD Operacional: valor actual del dato.
  + Horizonte de tiempo – 60-90 días
  + Puede o no tener clave con atributos de tiempo
  + Un dato puede ser cambiado (actualizado)
* Data Warehouse: fotografía del dato.
  + Horizonte de tiempo – 5-10 años.
  + Clave con atributos de tiempo
  + Una vez tomada la fotografía, el registro no se cambia.

4 – No volátil

* BD Operacional: los datos son actualizados regularmente registro a registro.
* Data Warehouse: el dato es cargado en el DW y accedido, pero una vez realizada la fotografía del dato, el dato en el DW no se modifica.

**Consecuencias de la no-volatilidad**

Manejo de datos mucho más simple.

* Elimina todo el proceso de actualización de datos en línea:
  + Backup y recuperación.
  + Transacción e integridad.
  + Detección y solución de deadlock.
* Es posible optimizar el acceso a datos.

**Data Mart**

* Base de datos para dar soporte a los procesos de decisión departamentales.
* Agrupa los datos asociados a un tema (ventas, compras, producción, distribución, etc.)

Ventajas

* Permite personalización local.
* Menor cantidad de datos históricos.
* Uso local de recursos de procesamiento.
* Los departamentos pueden elegir el software de análisis deseado.
* Menor costo unitario de procesamiento y almacenamiento.

Carga del Data Mart

Depende de la arquitectura usada:

* Alternativa 1: el DM es cargado desde el DW mediante programas de carga (Modelo de Inmon).
* Alternativa 2: el DM es una porción del DW y es cargado directamente desde las BD operacionales (Modelo de Kimball).
* Alternativa 3: solo existe el DM (no el DW) y el mismo es cargado directamente desde las BD operacionales. No es recomendable porque no permite integrar información de distintos DM para la toma de decisiones estratégicas, por ejemplo.

Metadato del DW y del DM

Metadato: directorio del dato

Función:

* Ayudar al analista a localizar los contenidos del DW o DM.
* Guiar el mapeo de datos, en la medida que el dato es transformado.
* Guiar los algoritmos usados para agregación/sumarización.

Metadato: información contenida

* Identificación de la fuente de los datos.
* Descripción de la transformación sufrida al pasar el dato al DW o DM.
* Información descriptiva del DM o DW (tablas, atributos, relaciones, etc.).
* Definición de los términos usados.

Soporte requerido para Crear y administrar el DW/DM

* Soporte para definir el ETL (programa de extracción, transformación y carga de datos).
* Soporte para administrar la BD.
* Soporte para administrar los metadatos.
* Soporte para acceso y análisis.

Almacenamiento de datos externos

* Se almacenan en el DW si es para toda la organización (evita redundancia, incompatibilidad, costos).
* Pueden ser personalizados al ingresar al DM o aplicación del usuario.
* Se debe guardar también el origen del dato (fuente, fecha de adquisición, cantidad adquirida, descripción, criterios de edición y filtrado aplicado, etc.).

La estructura de datos en el DW

*Agregado*: información agregada, importante para usuarios específicos

Metadatos

*Datos no agregados*: todos los datos que se van “tomando”.

*Datos en desuso:* datos viejos. No se eliminan, solo que se almacenan y quedan accesibles, pero puede que no sirvan para análisis de últimos años.

Niveles de agregación (ventas)

Ventas semanales por región 1995-2002

Ventas semanales por subproducto 1995-2002

Ventas detalladas 1995-1999

Ventas mensuales por país 1995-2002

Ventas detalladas 2000-2002

Flujo de los datos

Proceso de agregación

Proceso de archivado

Proceso de Extracción, Transformación y Carga

Población de Usuarios del DW

**Data Warehouse - Diseño, modelo y desarrollo**

Tareas iniciales del diseño de un DW

Pensar a un DM como un conjunto de Hechos Relacionados. No exhaustivo (no se van a tener cada una de las transacciones de las bases de datos operacionales; solo algunas de interés), no exclusivo (puede haber información repetida, debido al hecho de que la información está tomada de diferentes bases de datos.).

Construir una **Matriz: DMs (filas) x Dimensiones (columnas).**

* Listar de 10 a 30 DMs de fuente única.
* Listar las dimensiones (se obtienen mediante simple análisis).

Proyecto “El Fabricante”

Este ejemplo originaría un modelo del tipo “Estrella”.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data  Marts | Dimensiones | | | | | | | |
|  | **Tiempo** | **Sucursal** | **Articulo** | **Cajero** | **Cta. Cte.** | **Proveedor** | **Nivel Fidelidad** |
| **Ventas** | X | X | X | X | X |  | X |
| **Stock** | X | X | X |  |  |  |  |
| **Compras** | X | X | X |  |  | X |  |
| **Fabricación** | X | X | X |  |  |  |  |

Otras tareas iniciales del diseño de un DW

* Consensuar la matriz con los interesados.
* Conformar las dimensiones:
  + Una dimensión debe ser clara y debe definirse única para todos los hechos que usan la misma.

Hechos y dimensiones

Los *hechos* contienen los datos de estudio. Las *dimensiones* contienen metadatos (“clave ajena” que luego se va a propagar hacia los hechos) sobre dichos hechos.

Diseño de la tabla de Hechos

* Etapa 1: elegir un DM.
  + Fuente única.
  + Debe basarse en los datos disponibles (no en los ideales).
* Etapa 2: elegir el grano.
  + Definir el registro individual de la tabla.
* Etapa 3: definir las dimensiones.
  + Si una dimensión no coincide con el grano existen dos opciones:
    - Cambiar el grano (lo más común).
    - Descartar la dimensión.
* Etapa 4: definir los hechos.
  + Definir tantos hechos como sea posible en el contexto del grano elegido.

**Modelos de diseño de un DW**

**Modelo de Inmon**

Modelo: DMs dependiendo del EDW.

* Los datos originales están todos en el EDW.
* Los DMs son desarrollados a partir del EDW.
* EDW sirve a las funciones de la corporación.
* DMs sirven a las funciones de los departamentos (finanzas, ventas, compras, etc.).

FALTA IMAGEN “OLTP Data Sources”

Características

* EDW
  + Datos muy detallados.
  + Datos de toda la organización.
  + Historial de 1-10 años.
  + Diseñado para uso futuro no conocido.
  + Para toda la organización.
  + Data-driven (dirigido por los datos).
* DMs
  + Datos agregados.
  + Datos de interés del área.
  + Historial de 60-90 días.
  + Diseño para uso futuro concreto.
  + Para uso del departamento.
  + Requirement-driven (diseño dirigido por los requerimientos).

Objetivos del diseño del EDW

* Mejorar la velocidad de acceso a los datos.
* No reflejar requerimientos departamentales.

Estructura de datos

Estructura de datos del EDW: modelo E/R ligeramente desnormalizado.

* Se crea un lugar único para almacenar los datos a usar normal y frecuentemente juntos.
* Crear redundancia de datos a usar regularmente junto a otros datos. Ejemplo: un componente de producto aparece en la BOM, Inventario, MRP, etc.).
* Crear datos agregados cuando el dato es calculado y usado muchas veces. Ejemplo: el promedio de un alumno.
* Separar datos cuyo registro incluye campos de datos de uso dispar.

Objetivos de diseño de los DMs

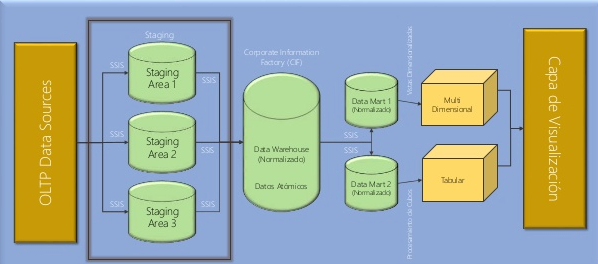
* Facilitar la comprensión de los datos.
* Reflejar requerimientos de departamentos (personas).

Estructura de datos de los DMs: modelo dimensional, estrella o copo de nieve.

**Modelo de Kimball**

DW

* Fuente consultable de datos de la empresa.
* DW = DM1 U (unión) DM2 U DM3 U… (no existe el DW como tal).
* Usar modelo dimensional.
* Con modelo E/R se pierde performance y comprensión.
* Propuesta:
  + Desarrollar modelos de datos global.
  + Implementar todo o por DM.



Componente de Integración de Datos

Propósito: mantener las vistas de los datos fuentes del DW. Realizar algoritmos apropiados.

Funciones del Componente de Integración de Datos:

* Extracción, transformación y carga (ETL).

Toma los datos de la fuente en su formato de origen, los transforma al formato del DW y los carga en el DW.

* + Carga inicial de datos al DW.
  + Carga de nuevos datos al DW
* Monitoreo de cambios en los datos.

Monitorea los datos en las fuentes e informa al ETL los cambios que considera de interés para el DW.

* + ETL: toma los datos de la fuente en su formato de origen, los transforma al formato del DW y los carga en el DW.

Mecanismos de monitoreo, dependiente del tipo de datos

* Cooperación: fuentes que poseen “triggers” u “otras capacidades” que permiten programar los cambios desde las mismas.
* Registro de Operación: poseen un “log” que puede ser inspeccionado para detectar cambios.
* Consultas SQL: permiten al componente de integración realizar consultas sobe la misma para detectar cambios de interés.
* Fotografías: se realizan lecturas periódicas de datos. Por comparación off-line de dos “fotos” sucesivas se detectan los cambios. Este mecanismo es el más visto en la actualidad.

**Evolución del DW**

La arquitectura del DW debería permitir manejar cambios en las fuentes de datos y en los requerimientos de los usuarios con el menor número de modificaciones de los componentes del DW.

**Modelos de desarrollo de DW/DMs**

* Top-down
  + Fuentes DW DMs (Inmon).
* Bottom-up
  + Fuentes DMs DW (Kimball).
* Paralelo
  + Fuentes DW DMs.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Top-down | Bottom- up | Hibrida | Federada |
| Autor | Inmon | Kimball | Diversos autores | Doug Hackney |
| Énfasis | EDW | DMs | EDW y DMs | Integrado a entornos BI heterogéneos. |
| Diseño | Modelo normalizado basado en la empresa. | Modelo dimensional del DMs usa esquema de estrella. | Modelos locales y uno o más esquemas de estrella. | Arquitectura de estructura; comparte dimensiones, hechos, reglas definiciones. |
| Data Set | EDW con datos atómicos y DMs con datos sumarizados. | Contiene datos atómicos y sumarizados. | Carga Datamarts con datos atómicos y sumarizados. | Uso de cualquier significado posible para integrar las necesidades. |
| Arquitectura | Varios niveles de áreas de interés y DMs dependientes. | Área de interés y datamarts. | Modelo empresarial normalizado y DMs iniciales. | Realidad del cambio en organizaciones y sistemas. |

**Modelado Dimensional**

Contiene la misma información que el E/R.

Diferencia: almacena los datos en una forma simétrica, cuyo objetivo de diseño es lograr:

* **Comprensión** por parte del usuario
* **Performance** para consultas
* **Flexibilidad** para cambiar

Definiciones

* **Atributos:** generalmente describen una característica de una cosa tangible (tipo de producto, marca, empaque). Se organizan o agrupan en **dimensiones**.
* **Facts (Hechos):** es algo que se desea observar.
  + Numéricos
  + Textos

Clasificación de hechos numéricos

* **Aditivos:** se pueden sumar en todas las dimensiones (productos vendidos por mes, productos de una familia de productos).
* **Semi-aditivos:** se pueden sumar en todas las dimensiones excepto en la dimensión tiempo. En la dimensión tiempo se suman y se divide por el número de períodos de tiempo.
* **No Aditivos:** medidas de intensidad no aditivas en todas las dimensiones. Se usa la función SQL\_AVG.

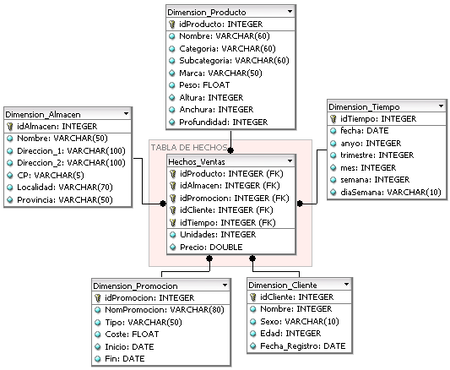
Modelado dimensional de hechos

Consta de:

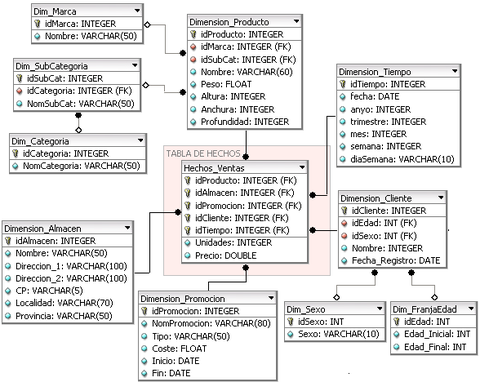
* **Tablas dimensionales**: conjunto de atributos. Posee, en general, una sola clave primaria (PK).
* **Tabla de hechos**: conjunto de hechos. Posee muchas claves foráneas (FK).

Modelos dimensionales

* Esquema estrella



* + Ejemplo:
    - Tabla de Hechos: Ventas.
    - Tablas de Dimensiones: Almacén, Producto, Tiempo, Cliente, Promoción.
  + Ventaja: para cualquier tipo de consulta, el máximo recorrido es un join de dos tablas (hecho + dimensión).
* Esquema copo de nieve [No Recomendado]



* + Ejemplo:
    - Tabla de Hechos: Ventas.
    - Tablas de dimensiones: Marca, Producto, Subcategoría, Categoría, Tiempo, Cliente, FranjaEdad, Sexo, Promoción, Almacén.
  + Desventaja: las consultas son mucho más lentas/complejas.

Dimensiones Conformadas

* Son dimensiones comunes a toda tabla de hechos que utiliza esa dimensión. Ejemplo: la dimensión tiempo se define de modo que sea común a toda tabla de hechos que la utiliza.
* Grano (grain): implica definir el registro individual de una tabla de hechos. Ejemplo: venta (unidades o $) por factura, venta diaria de un producto en un local.

Clasificación de Hechos

* Hecho base: es el hecho almacenado en la tabla de hechos.
* Hecho derivado:
  + Aditivo: puede ser generado desde los otros hechos de una misma tabla. Es decir, comparte el mismo grano. Se puede mostrar al usuario como perteneciendo a dicha tabla.
  + No aditivos: un valor acumulado que es expresado en diferente grano. No puede mostrarse en la misma tabla de hechos.

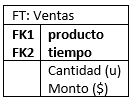
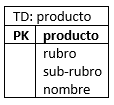
Hecho Agregado

Es el almacenamiento de sumarizaciones de un hecho con el objeto de mejorar la performance de las consultas.

Para ello se definen:

* Jerarquías de agregación
  + Niveles de agregación.

Ejemplo:



* + Jerarquía 1: TD tiempo
    - Agregación
      * Nivel 1: mes
      * Nivel 2: año
  + Jerarquía 2: TD Producto.
    - Agregación:
      * Nivel 1: sub-rubro
      * Nivel 2: rubro
  + Jerarquía 3: TD Tiempo, TD Producto.
    - Agregación:
      * Nivel 1: mes, sub-rubro
      * Nivel 2: año, sub-rubro
      * Nivel 3: mes, rubro
      * Nivel 4: año, rubro

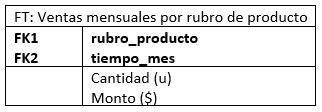
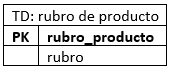
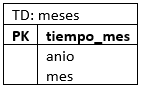
Jerarquías de Agregación

Sea F una tabla de hechos y D un conjunto de tablas dimensionales, se define el conjunto de jerarquías de agregación H, donde h H es una jerarquía de agregación de F respecto a un subconjunto de d D y tiene un conjunto ordenado finito de niveles de agregación N(h).

* Todo h H son jerarquías distintas.
* Todo n N(h) tiene el mismo conjunto de d que h.
* Cada n N(h) satisface un criterio de crecimiento agregando una o más d respecto al n precedente y manteniendo las restantes d no agregadas.
* Cada n N(h) tiene una nueva F generada des de la F del nivel precedente siguiendo las agregaciones definidas por n N(h)
* Cada h H define su primer n desde la F primitiva.

Tablas Agregadas

* Por cada nivel de una jerarquía de agregación se puede generar una tabla de hechos que se denomina agregada (se suele generar cuando se tienen muchas consultas sobre ese hecho agregado).
* Tablas de hechos agregadas:
  + Se decide generarlas a partir de las estadísticas de uso de los datos por parte de los usuarios.
  + Reducen el tiempo de respuesta a las consultas de los usuarios.
  + Se pueden eliminar cuando las mismas dejan de ser consultadas.
* Ejemplo de tabla agregada:



* + Jerarquía 3: TD tiempo, TD Producto
    - Agregación:
      * Nivel 3: mes, rubro.

**Tecnologías OLAP (On-Line Analytical Processing)**

Tecnología de información diseñada para permitir acceso y visualización ad-hoc de datos.

Son aplicaciones usadas por analistas y gerentes quienes frecuentemente quieren vistas muy agregada de los datos.

* Por ejemplo: ventas mensuales por productos y por regiones.

Las bases de datos OLAP generalmente son actualizadas en batch y a veces desde múltiples fuentes. A su vez, son optimizadas para análisis.

OLAP es sinónimo de *vista multidimensional* de datos. Para que un dato pueda ser visto desde múltiples dimensiones, es necesario que sea *multidimensional*.

Ejemplo dato NO multidimensional – Consolidación en BD relacional

|  |  |  |
| --- | --- | --- |
| **Producto** | **Región** | **Ventas** |
| A  A  A | Este  Oeste  Central | 50  60  100 |
| **A** | **Total** | **210** |
| B  B  B | Este  Oeste  Central | 40  70  80 |
| **B** | **Total** | **190** |
| C  C  C | Este  Oeste  Central | 90  120  140 |
| **C** | **Total** | **350** |
| D  D D | Este  Oeste  Central | 20  10  30 |
| **D** | **Total** | **60** |
| **Total** | **Este** | **200** |
| **Total** | **Oeste** | **260** |
| **Total** | **Central** | **350** |
| **Total** | **Total** | **810** |

La tabla relacional incluye totales de ventas por producto y por región pre-calculados (consolidación).

Esto evita tener que calcular los totales en el momento en que se los requiere.

El resultado es: consistencia y menor tiempo de respuesta.

Ejemplo dato multidimensional

|  |  |  |
| --- | --- | --- |
| **Producto** | **Región** | **Ventas** |
| A  A  A  B  B  B  C  C  C  D  D  D | Este  Oeste  Central  Este  Oeste  Central  Este  Oeste  Central  Este  Oeste  Central | 50  60  100  40  70  80  90  120  140  20  10  30 |

*Hecho* a observar: VENTAS.

*Dimensiones* que caracterizan hecho:

PRODUCTO [nombre]

DISTRIBUCION GEOGRAFICA [región]

Una matriz bi-dimensional es más clara que la tabla relacional previa.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Este** | **Oeste** | **Central** |
| **A** | 50 | 60 | 100 |
| **B** | 40 | 70 | 80 |
| **C** | 90 | 120 | 140 |
| **D** | 20 | 10 | 30 |

Consolidación en BD multidimensional

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Este** | **Oeste** | **Central** | **Total** |
| **A** | 50 | 60 | 100 | **210** |
| **B** | 40 | 70 | 80 | **190** |
| **C** | 90 | 120 | 140 | **360** |
| **D** | 20 | 10 | 30 | **60** |
| **Total** | **200** | **260** | **350** | **810** |

Si a este ejemplo se le agregara una dimensión más de tiempo, podría ser un cubo.

La consolidación en una BD multidimensional es simple de visualizar.

Jerarquías y Dimensiones

Región Total

Oeste

Central

Este

Total Producto

C

D

B

A

* En una jerarquía simple, cada hijo tiene un solo padre.

Región Total

Oeste

Central

Este

Illinois

Michigan

Ohio

Chicago

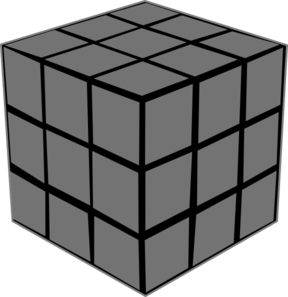
Champaign

* Una jerarquía simple puede tener muchos niveles.
* Desde el top se puede realizar un “drill down” a sucesivos niveles de detalles.

Dimensiones con jerarquías

* Las dimensiones con jerarquías permiten a las ciudades, estados y regiones existir en una sola dimensión.

Este



Producto

Tiempo

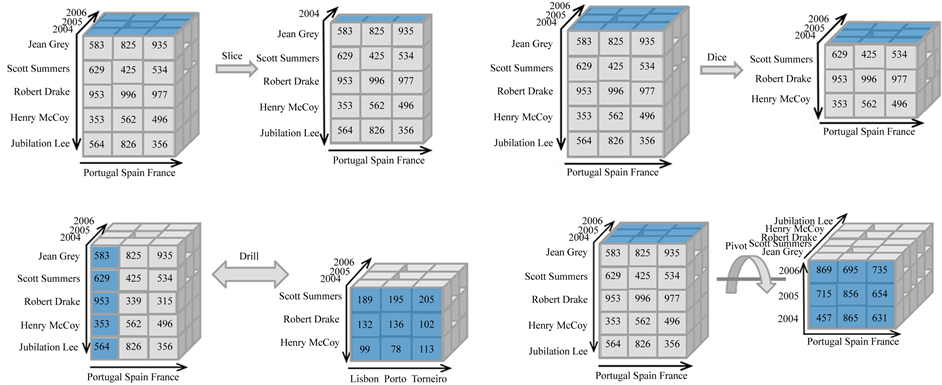
* New York.
* New Jersey.
* Pennsylvania.

Central

* Ohio.
* Michigan.
* Illinois.
  + Chicago.
  + Champaign.

Slice and dice

Un arreglo n-dimensional tiene n(n-1) vistas. La posibilidad de “rotar el cubo de datos” es la técnica principal para reportes multidimensionales y es llamada “slice and dice”.



Tipo de dato: Time-Series

Es un *metadato* que indica cuál es la información que está almacenada en la celda.

Almacenando una serie de tiempo completa en cada celda elimina la necesidad de tener una dimensión tiempo separada. A modo de ejemplo, tener en una celda tengo todas las ventas del mes de enero y en otra todas las de febrero.

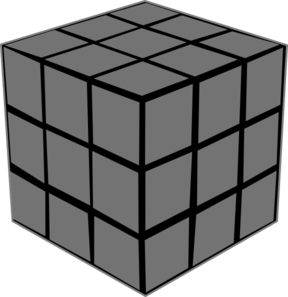
|  |  |  |  |
| --- | --- | --- | --- |
| Tiendas | xxxx | xxxx | xxxx |
| xxxx | xxxx | xxxx |
| xxxx | xxxx | xxxx |
|  | Productos | | |

Sin el tipo de dato Time-Series es necesario definir una dimensión *Tiempo.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Productos | Enero | Febrero | Marzo | …. | Diciembre |
|  |  |  |  |  |
|  |  |  |  |  |

Ejemplo: por cada celda que se tiene, se “pega” toda esa información como metadato.

* Start date = 1/1/94
* Periodicity = Daily, business days only.
* Conversion = Summation
* Long description = Variable=sales, Product=Nuts, Region=East.
* Data type = Numeric, single precisión.
* Calendar = 445 Fiscal year.



Producto

Tienda

Días

Tecnología de CUBOS

Es la tecnología que utilizan los sistemas OLAP para almacenar los datos y permitir la vista multidimensional de los mismos.

*Fortalezas:*

* La performance es *óptima* para dar soporte a procesos de decisión.
* Puede ser optimizada para el acceso rápido a los datos.
* Si se conoce un patrón de uso de los datos esta tecnología es la mejor.
* Permite usar fácil “sliced and diced” (rotación del cubo).
* Permite examinar los datos desde muchas dimensiones.
* No es necesario escribir SQL.

*Debilidades:*

* Maneja reducida cantidad de datos.
* El tiempo de carga puede ser grande.
* La estructura de datos no es flexible, solo se puede realizar accesos respondiendo al diseño.
* El soporte para vínculos dinámicos (activar procesos batch ante ciertos eventos) es cuestionable (no siempre funciona bien).

Beneficios de la tecnología OLAP para acceder al DW

*Beneficios para el usuario:*

* Aísla al usuario del lenguaje SQL y del modelo de datos
* Mejora la performance de las consultas.
* Aumenta la capacidad de cálculos de información de negocio.
* Soporta un amplio rango de aplicaciones (se pueden armar para diferentes tipos de dominios).

*Beneficios para responsables de sistema:*

* Facilita la administración del sistema.
* Automatiza el mantenimiento de índices y sumarizaciones.
* Reduce la carga sobre el DW.
* Libera al personal de sistemas de generación de reportes.
* Permite un control sobre los datos para análisis.

Características deseables para una aplicación OLAP

* Tipo de dato: Time-Series.
* Dimensiones jerárquicas (simple y múltiples niveles).
* Variables derivadas (Utilidad bruta = Ventas – Costos).
* Datos repetitivos (sparse data).
* Slice and dice (rotación del cubo).

Variantes de los soprtes OLAP

* ROLAP (Relational On-Line Analytical Processing)
  + Genera los cubos multidimensionales dinámicamente al momento de realizar las consultas. Almacena los datos en estructuras relacionales.
* MOLAP (Multidimentional On-Line Analytical Processing)
  + Almacena físicamente los datos en estructuras multidimensionales. Dispone de estructuras de almacenamiento específicas y técnicas de compactación de datos.
* HOLAP (Hybrid On-Line Analytical Processing)
  + Constituye un sistema híbrido entre MOLAP y ROLAP. Los datos agregados y precalculados se almacenan en estructuras multidimensionales y los de menor nivel de detalle en estructuras relacionales.

Soporte Inteligente para Negocios

Base de datos operacional

Base de datos operacional

Base de datos operacional

Warehouse

Data Mining

OLAP Server

Soporte para Data Mining

Soporte Inteligencia de Negocios

**Data Mining**

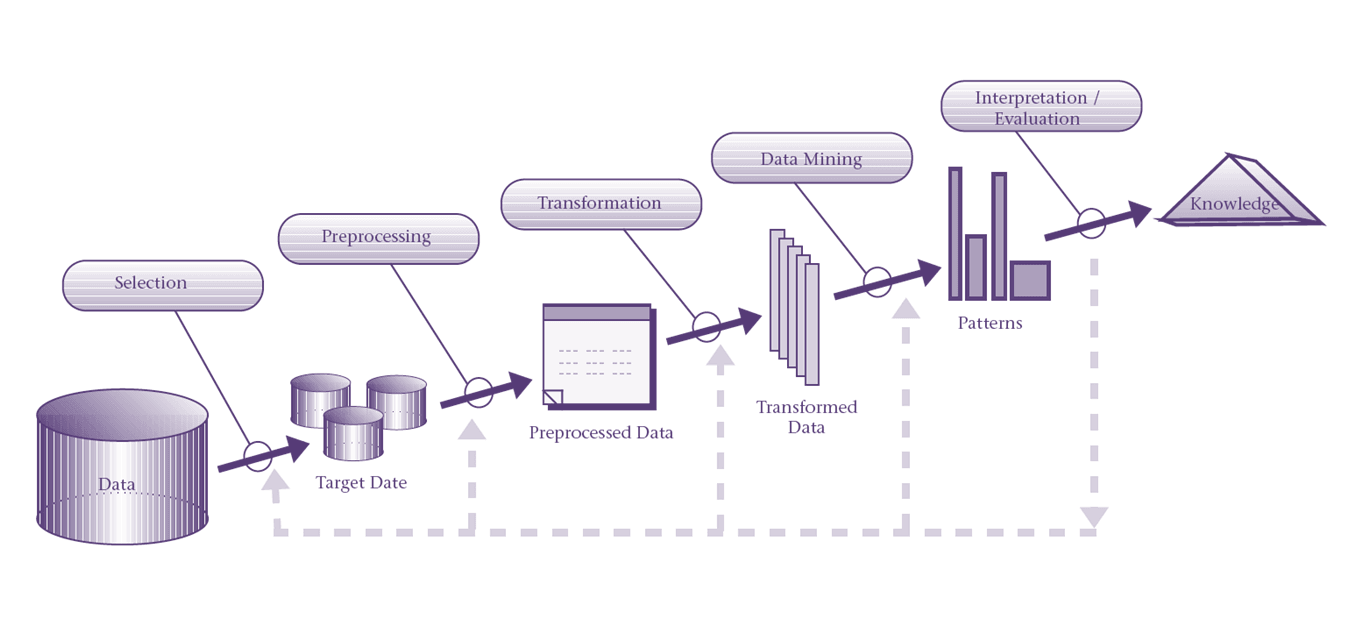
¿Qué es?

*“Extracción no trivial de información implícita, previamente no conocida y potencialmente útil desde los datos.”*

Definiciones:

* Variedad de técnicas para identificar información o conocimiento desde un conjunto de datos, y extraer éstos de modo tal que puedan ser dispuestos para su uso en diversas áreas.
* Los datos son frecuentemente voluminosos, pero en su forma actual no se puede hacer uso directo de ellos. Es la información oculta en ellos la que es útil.
* Es un proceso analítico diseñado para explorar grandes cantidades de datos en búsqueda de patrones consistentes y/o relaciones sistemáticas entre variables, y luego validar el hallazgo aplicando los patrones encontrados a nuevos conjuntos de datos.

Proceso de Data Mining



BD para Mining -> (Selección) -> Datos objetivos -> (Pre-Proceso) -> Datos procesados -> (Transformación) -> Datos transformados -> (Data Mining) -> Patrones -> (Interpretación) -> Conocimiento.

1º Selección

* Selección del conjunto de datos a utilizar según los objetivos del proceso de Mining a realizar.
* Cuando se tiene un exceso de registros de datos (población muy grande) se deben utilizar técnicas estadísticas para elegir el conjunto de registros representativos (muestra) a procesar.
* Técnicas estadísticas: diseño de experimentos.
* Seleccionar el conjunto de datos a usar:
  + Si hay exceso de registros de datos
    - Generar dos conjuntos de registros representativos:
      * Entrenamiento.
      * Validación.

2º Preproceso

Los datos deben ser analizados a efectos de determinar valores erróneos, valores incorrectos o datos faltantes. Estrategia a seguir:

* Descartarlos.
* Omitir el registro.
* Inferir valores a partir de valores conocidos.
* Tratarlo como un valor especial.
* Usar un valor promedio.

3º Transformación

Los datos pueden requerir transformaciones por diferentes razones:

1. Debido a las diferencias de escalas de los mismos, a efectos de evitar problemas numéricos durante el procesamiento.
   1. Estrategia: realizar transformación definiendo una escala apropiada (generalmente 0-1).
2. Debido a la presencia de valores cualitativos (no numéricos), que requieren ser cuantificados mediante escalas apropiadas.

4º Data mining

Procesamiento de los datos mediante técnicas apropiada a efectos de encontrar (si existen) patrones.

5º Interpretación

Interpretar los patrones de comportamiento en el dominio del negocio a efectos de generar información de utilidad (conocimiento). Interpretar es validar los patrones obtenidos con otro conjunto de datos.

Comparación de DM y consultas de BD

Las consultas de una DB pueden responder a:

* Ventas del último mes de un producto.
* Ventas agrupadas por edades de los clientes.
* Listado de clientes que no renuevan la póliza de seguro.

Data Mining – Infiere conocimiento desde los datos, ayudando a responder preguntas como:

* ¿Qué características comparten los clientes que no renovaron la póliza de seguros y cómo difieren de aquellos que renovaron la póliza de seguro?
* ¿Por qué la sucursal Buenos Aires es más rentable?

Modelos Data Mining

* Modelos de verificación
  + Toma una hipótesis definida por el usuario y realiza un test de validación.
* Modelos de descubrimiento
  + El sistema automáticamente descubre patrones de comportamiento de los datos.

Inconvenientes en Data Mining

* *Información limitada:* las bases de datos no son diseñadas para tareas de DM.
* *Valores erróneos:* errores en los atributos o en su definición.
* *Valores incorrectos o faltantes.*
* *Incertidumbre:* refiere a la incertidumbre respecto a la veracidad del dato.
* *Actualización:* las bases de datos son dinámicas y se deben mantener actualizadas las reglas.
* *Relevancia:* importancia de un campo de dato para el DM, por ejemplo: CP.

Potenciales aplicaciones

* Ventas/mercado
  + Identificar patrones de compra de los consumidores.
  + Asociaciones entre consumidores y características geográficas.
  + Análisis de canastas de mercado.
* Bancarias
  + Detectar patrones de uso fraudulento de tarjetas de crédito.
  + Identificar clientes fieles.
  + Gasto de tarjetas de crédito por grupos de clientes.
  + Encontrar correlaciones ocultas entre indicadores financieros.
* Seguros
  + Identificar patrones de conducta de clientes de riesgo.
  + Identificar comportamientos fraudulentos.
* Transporte
  + Identificar patrones de carga.
* Medicina
  + Identificar terapias exitosas para diferentes enfermedades.
  + Identificar grupos potenciales de adquirir determinadas enfermedades.

Funciones de Data Mining

* *Clasificación.*
  + Los sistemas de DM “aprenden” a clasificar los datos, formulando reglas de clasificación.
  + Ejemplo: base de datos de clientes de un banco.
    - Pregunta: ¿el nuevo solicitante de crédito, es una buena inversión?
    - Regla típica formulada:
      * SI ESTADO = casado AND INGRESO > 30000 AND PROPIETARIO DE VIVIENDA = si

ENTONCES TIPO\_INVERSION = buena

* *Asociación.*
  + Regla que asocia atributos de una relación a otros.
  + Las estrategias orientadas a conjuntos son los medios más eficientes para descubrir tales reglas.
  + Base: un conjunto de ítems y un conjunto de registros.
  + Ejemplo: base de datos de un supermercado:
    - El 40% del total de clientes compara A y C, de ellos el 75% también compran B, D y E.
    - Regla: A, C --> B, D, E
      * Soporte = 40 % y Factor de confianza = 75%.
* *Clustering/Segmentación.*
  + Partición de una base de datos en grupos similares según algún criterio o métrica.
  + Clustering por similaridad es usado en muchas disciplinas científicas.
  + Ejemplos: clustering de moléculas, segmentación de una base de datos de clientes.
* *Patrones Secuenciales/Temporales.*
  + Analizan conjuntos de registros relacionados y detectan patrones que ocurren durante un período de tiempo.
  + La diferencia entre las reglas de secuencia y otras reglas es el factor temporal.
  + Ejemplo: base de datos de un comercio:
    - Puede ser usada para descubrir el conjunto de compras que preceden la compra de un microondas.

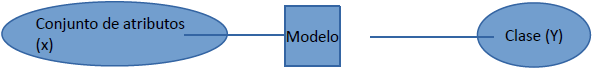
Técnicas y herramientas de uso frecuente en Data Mining

* Visualización.
  + Permite el análisis exploratorio de los datos.
  + Hace más visibles los patrones (OLAP, diagramas 2D, 3D, etc.)
* Análisis de Clusters.
  + Define grupos de una población sobre cada uno de los cuales se pueden realizar nuevos análisis.
  + Clustering Jerárquicos.
  + K-Clusters.
* Inducción
  + Árboles de decisión.
    - Usados para clasificar objetos.
    - Se construye usando un *conjunto de datos de entrenamiento.*
    - Problema: muchas ramas.
  + Basados en reglas.
    - La forma de representación más común.
    - Son simples e intuitivas.
    - No estructuradas y poco rígidas.
    - Problema:
      * Difíciles de mantener.
      * No apropiadas para todos los dominios.
    - Formato:
      * Si X entonces Y.
    - Tipos
      * Proposicionales.
      * Asociación.
* Redes Neuronales.
  + Modelo matemático complejo que puede capturar el patrón de comportamiento de un conjunto de datos.
  + Permite realizar interpolaciones.
  + Problemas:
    - La red resultante es vista como una caja negra.
    - No explica los resultados, por lo cual es difícil para los usuarios interpretar los resultados.
    - No permite la intervención del usuario.
* Técnicas Estadísticas.
  + Se pueden usar en varias etapas de un proceso de DM.
  + Limpieza de los datos.
    - Para remover datos erróneos o irrelevantes, para generar datos faltantes.
  + Análisis exploratorio de los datos.
    - Frecuencia, histogramas, etc.
  + Selección de datos.
    - Para elegir un conjunto representativo (muestra).
  + Análisis de datos.
    - Define medidas de asociación y relación entre datos.

**Clasificación**

Objetivo

Asignar objetos a una de varias clases definidas:



Registro (X, Y)

* X: conjunto de atributos.
  + Discretos o continuos.
* Y: atributo objetivo, clase, grupo.
  + (Discreto) diferencia con regresión.

*“Tarea de aprender una función que mapea cada conjunto de atributos Xa una clase predefinida Y”*

Uso

* *Modelo descriptivo:* permite describir características de una clase **y**, o las características que diferencian dos clases.
* *Modelo predictivo:* para predecir la clase **y** a la que pertenece un objeto desconocido.

Proceso general para generar un modelo de clasificación

*Algoritmo de aprendizaje*

Para generar un modelo que mejor “relaciona” **X con Y**

Conjunto de datos entrenamiento Modelo de aprendizaje (inducción) Modelo Deducción Conjunto test Modelo Validado

*Validación del modelo – métricas*

* Exactitud: cantidad predicha correcta / total de predicciones.
* Tasa de error: cantidad predicha incorrecta / total de predicciones.
* Precisión: objetos de la “clase” (buenos pagadores de verdad, por ejemplo) recuperados / total de objetos recuperados como “clase” (los que se clasificaron como buenos pagadores, por ejemplo).
* Recall: objetos de la “clase” recuperados / total de objetos de la “clase” (total exacto conocido de buenos pagadores del conjunto de datos).

*Técnicas de clasificación*

* Árboles de clasificación.
* Clasificación en base a reglas.
* Análisis discriminante.
* Redes neuronales.
* Máquinas de soporte vectorial.
* Redes de Bayes.

*Árboles de clasificación*

* Nodo raíz: no tiene entradas y tiene una o más salidas.
* Nodo interno: tiene una sola entrada y dos o más salidas.
* Nodo final: tiene una sola entrada y no tiene salidas. Los nodos finales representan las clases.
* Cada nodo final es asignado a una clase.
* Los nodos raíz e internos contienen condiciones de test de atributos (CTA) que permiten dividir el conjunto de registros en subconjuntos con diferentes características.

*Construcción de un árbol de clasificación (inducción)*

Características:

* Número exponencial de alternativas.
* Algunos son más exactos que otros.
* No es posible asegurar el óptimo.

**Algoritmo de Hunt (recursivo)**



Excepciones

Si algún nodo hijo resulta vacío (sin registro) el nodo se declara al final y se lo asigna a la clase mayoritaria de los registros del nodo padre.

Si en la etapa 2 todo registro asociado con Dt tiene idéntico valor para sus atributos excepto para la clase, no es posible separar estos registros. El nodo es declarado final y asignado a la clase de la mayoría de los registros de dicho nodo.

Consideraciones de diseño

División del conjunto de registros:

Cada etapa del proceso (inductivo) debe elegir un atributo para la CTA. Para ello, el algoritmo debe proveer un método para especificar la CTA para diferentes tipos de atributos y una medida objetiva para evaluar las bondades de cada CTA.

Condición de parada:

* Expandir cada noto hasta que todos los registros pertenezcan a la misma clase o todos tengan idéntico valor de los atributos.
* Terminación temprana para evitar sobre entrenamiento.

Métodos para la CTA

* *Atributo binario:* genera dos posibles resultados.
* *Atributo nominal:* puede tener varios valores.
  + Dimensión múltiple.
  + División binaria sucesiva.
* *Atributo ordinal:* puede usarse división múltiple o binaria (en este caso mantener la relación de procedencia).
* *Atributo continuo:* puede ser expresado como un test de comparación con
  + División binaria: (A < y) o (A >= v)
  + División múltiple: vi <= A + l, i = l, .., .. k

{N}

{S}

{L, I}

{L, I}

{S, N}

**Errores de un modelo**

*Errores de entrenamiento:* es producido por los registros del conjunto de entrenamiento que resultan mal clasificados.

Fórmula:

**

Error de entrenamiento

* tl: nodo terminal.
* k: cantidad de nodos terminales.
* e(ti): registros de entrenamiento mal clasificados en ti.
* n(ti): total de registros de entrenamiento clasificados en ti.

Error de generalización

Es producido por los registros del conjunto de validación que resultan mal clasificados. Un buen modelo debe tener bajo error de entrenamiento y de generalización.

Sobre-entrenamiento

Cuando un modelo tiene un error de entrenamiento muy bajo y un alto error de generalización.

* Causas posibles:
  + Presencia de ruidos (errores en los registros).
  + Falta de registros representativos en el conjunto de entrenamiento.

**Clasificación en base a reglas**

Técnica de clasificación de objetos en base a un conjunto R de reglas “Si… entonces…”

Modelo de clasificación: R = (r1 v r2 v … rk)

R1 = (condición) -> y1

Condición1 = (A1 op v1) and (A2 op v2) and … (Ak op vk).

Donde:

* Ak = atributo.
* vk = valor.
* op = operador elegido de { =, <>, <=, >=, <, > }

(Ak op vk) se denomina conjunto k de la precondición.

Calidad de una regla de clasificación

Dado un conjunto de registros D y una regla de clasificación r: A -> y

Se define:

* Cobertura (r) = |A| / |D|
* Exactitud (r) = |A and y| / |A|

Donde:

* |A| = cantidad de registros de D que satisface la precondición.
* |A and y| = cantidad de registros de D que satisface la precondición y el consecuente.
* |D| = cantidad total de registros.

*Ejemplo*

Sobre un total de 100 registros de clientes: 40 clientes compran A y C, de los cuales 30 también compran B, D y E

* Regla: A, C -> B, D, E
* Cobertura (r) = |A| / |D| = 40 / 100 = 0,4.
* Exactitud (r) = |A and y| / |A| = 30 / 40 = 0,75.
* Confidencia 0 75% y el soporte = 40%.

Propiedad de un conjunto de reglas R

R es **mutuamente exclusivo** si como máximo se dispara una r para un registro.

R es **exhaustivo** si posee una r para cada combinación de valores de los atributos -> todo registro es cubierto por al menos una r.

*Ejemplo*

Asumiendo que (Tº corporal) y (parto) son variables binarias.

* R = r1 (Tº corporal = frio) -> NO Mamífero.
* R2 = (Tº corporal = caliente) (parto = sí) -> Mamífero.
* R = r1 (Tº corporal = caliente) (parto = no) -> NO Mamífero.
* R es mutuamente exclusivo y exhaustivo.

Puede ocurrir que R

* *No exhaustivo:* un registro no tiene cobertura.
  + Se agrega una regla por default rd: () -> yd para cubrir los casos restantes.
  + rd se dispara si todas las otras r fallan.
  + yd se asigna a la clase mayoritaria de los registros cubiertos.
* *No mutuamente excluyente:* un registro es cubierto por 2 o más reglas con resultado distinto.
  + Solución
    - Reglas ordenadas: por prioridad decreciente (de exactitud, cobertura, etc.).
    - Reglas no ordenadas: permite que un registro dispare varias reglas.

**Big Data**

Big Data es la tendencia en el avance de la tecnología que ha abierto las puertas para un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costos cargarlos a tecnologías de bases de datos tradicionales para su análisis.

Big Data aplica para toda aquella información que no puede ser procesada o analizara utilizando procesos o herramientas tradicionales.

¿Data Warehouse vs Big Data? -> En la actualidad, la respuesta sería “No hay un reemplazo de una por la otra”.

Data Warehouse

* Proceso complejo y costoso.
* Datos estructurados.
* Datos que se sabe de ante mano que tienen un alto valor de negocio.

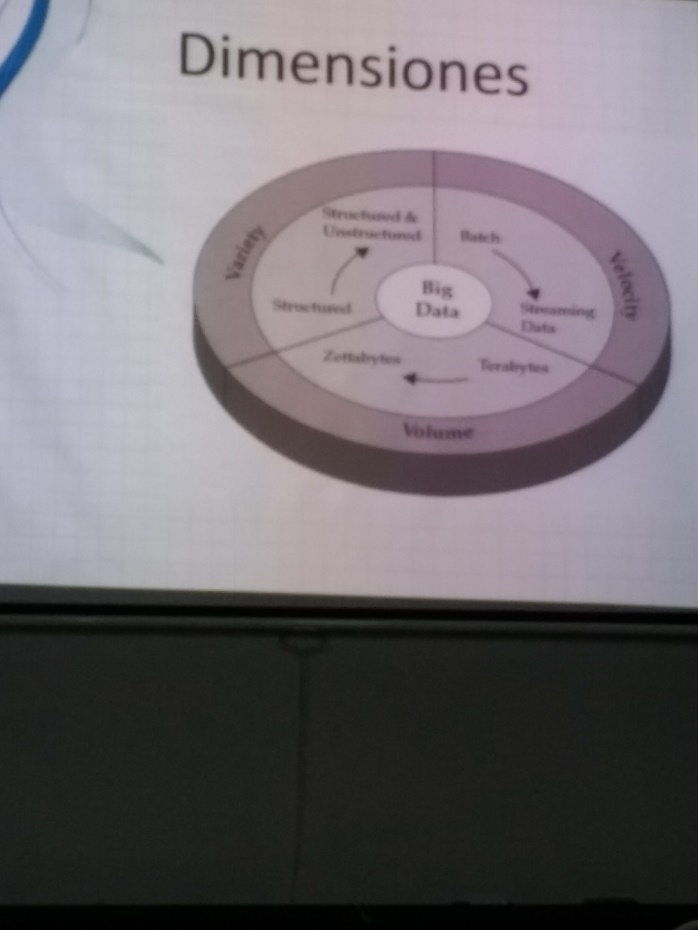
Big Data

* Proceso más simple y menos costoso.
* Datos estructurados, no estructurados y semi-estructurados.
* No se tiene certidumbre de ante mano sobre la calidad y valor de los datos.

Usos de Big Data

* Las empresas generan gran cantidad de información, pero al no poder procesarla, no pueden establecer su valor, y se pierden oportunidades de negocio.
* Existen casos que, para analizar determinados datos, el hecho de trabajar sobre muestras y no sobre el conjunto total hace que se pierda calidad en el análisis.
* Análisis de datos en procesos iterativos o exploratorias.

Dimensiones de Big Data



* Volumen.
  + Se habla de términos de Petabytes y exabytes de datos.
* Variedad
  + Datos en dispositivos móviles, audio, video, sistema GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc.
* Velocidad.
  + Datos que requieren que la velocidad de respuesta sea lo demasiado rápida para lograr obtener la información correcta en el momento preciso.

Origines de datos

* Web and Social Media.
* Machine-to-Machine.
* Big Transaction Data.
* Biometrics.

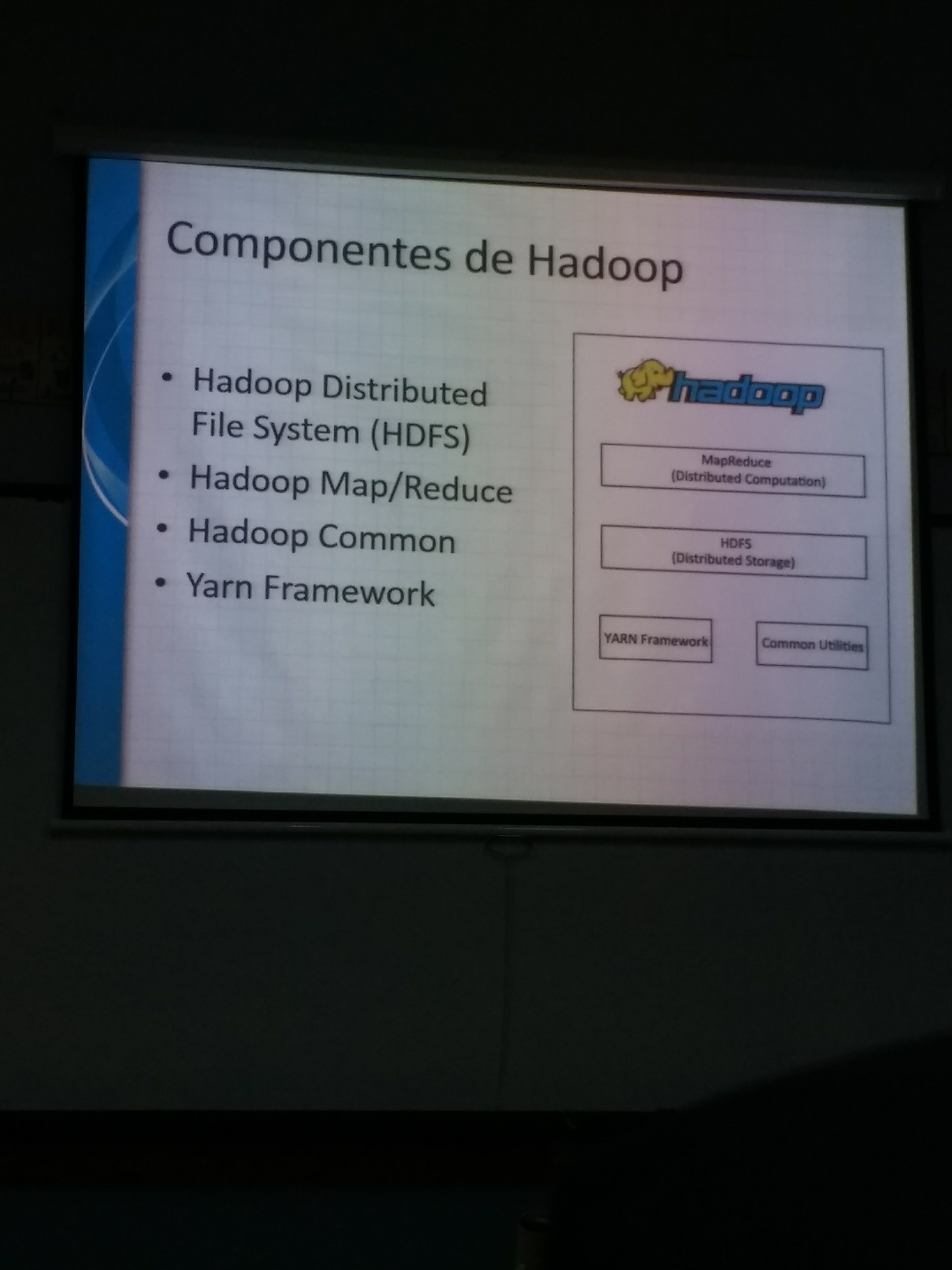
Plataforma de Big Data

* ***Hadoop.***
* Avro.
* Cassandra.
* Chukwa.
* Flume.
* Hbase.
* Hive.
* JAQL.
* Lucene.
* Oozie.
* Pig.
* Zookeeper.

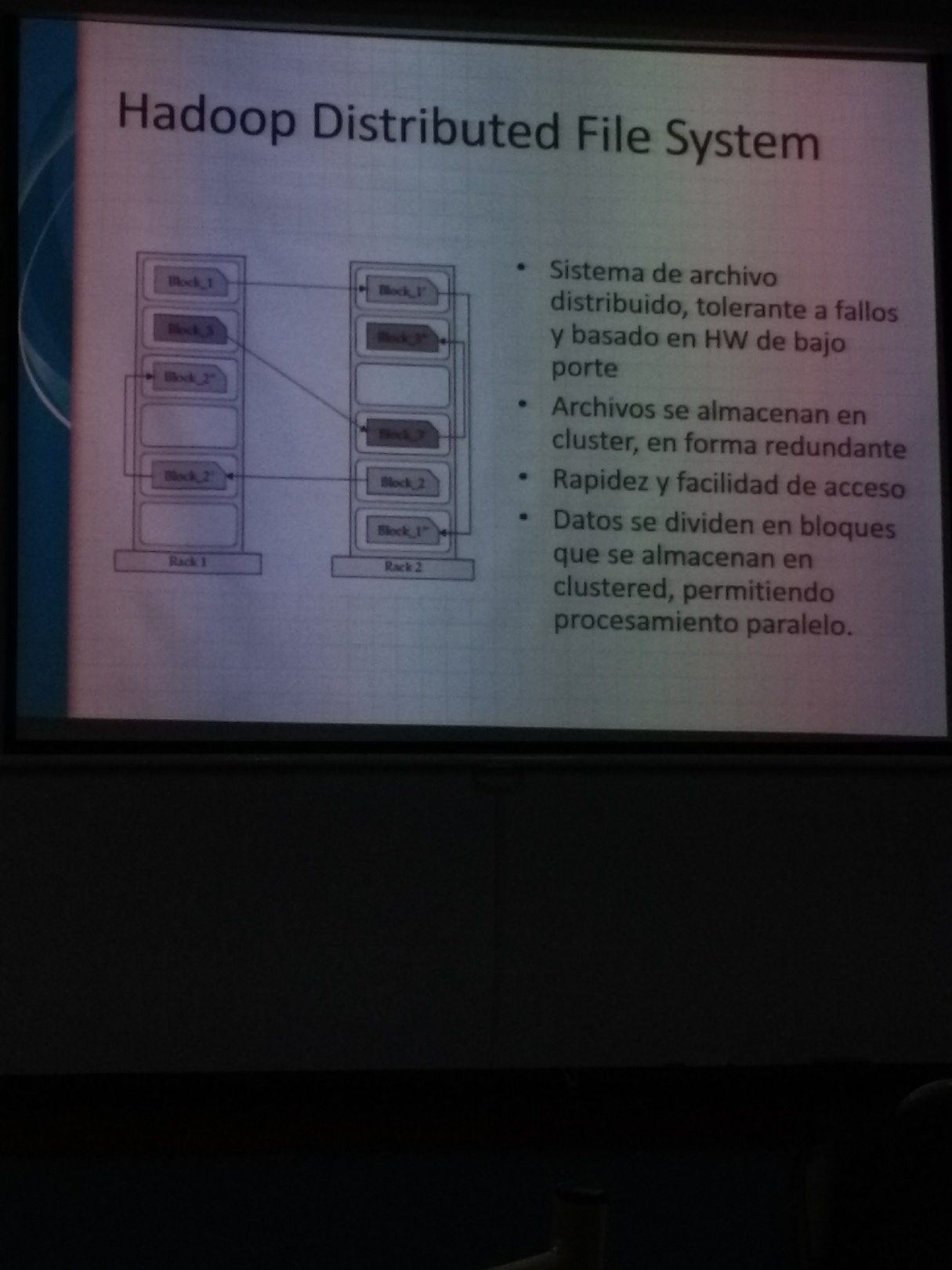
**Hadoop**

* Proyecto de Apache (Open Source).
* Diseñado para tener grandes capacidades de escalabilidad.
* Maneja y recuperación automática de errores.
* Basado en la utilización de hardware de bajo porte.
* Reducción de costos de hardware, software y mantenimiento.

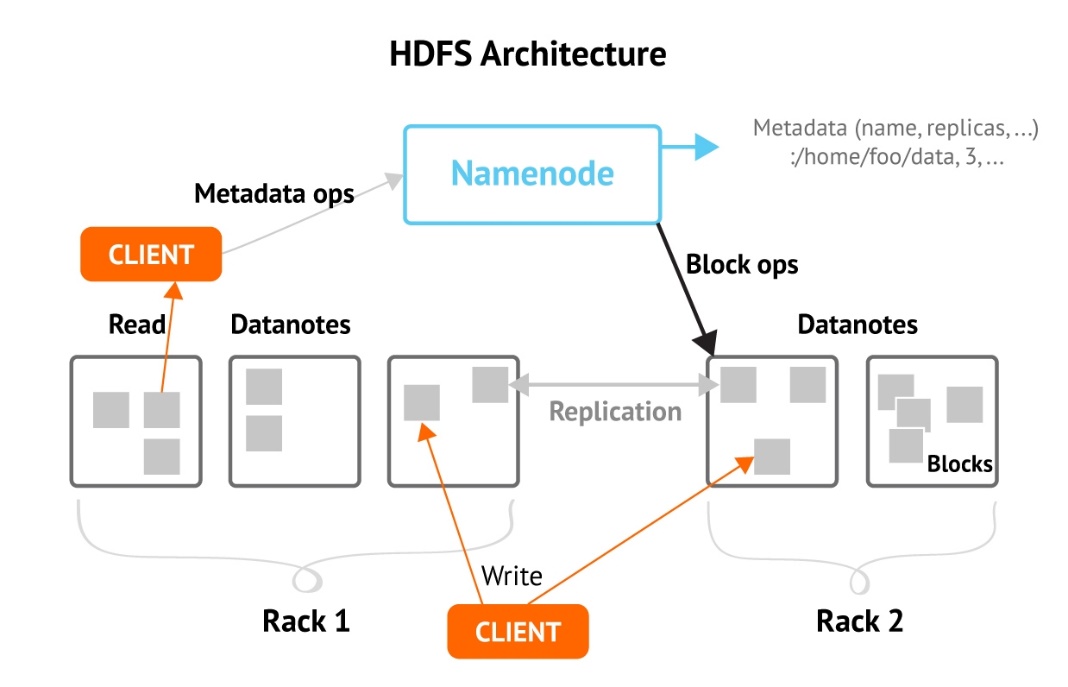
*Componentes de Hadoop*



* *Hadoop Distributed File Systema* (HDFS) (sistema de archivos).
* *Hadoop Map/Reduce* (sistema/lógica de procesamiento).
* Hadoop Common (librerías comunes).
* Yarn Framewrok (herramientas de programación de tareas).

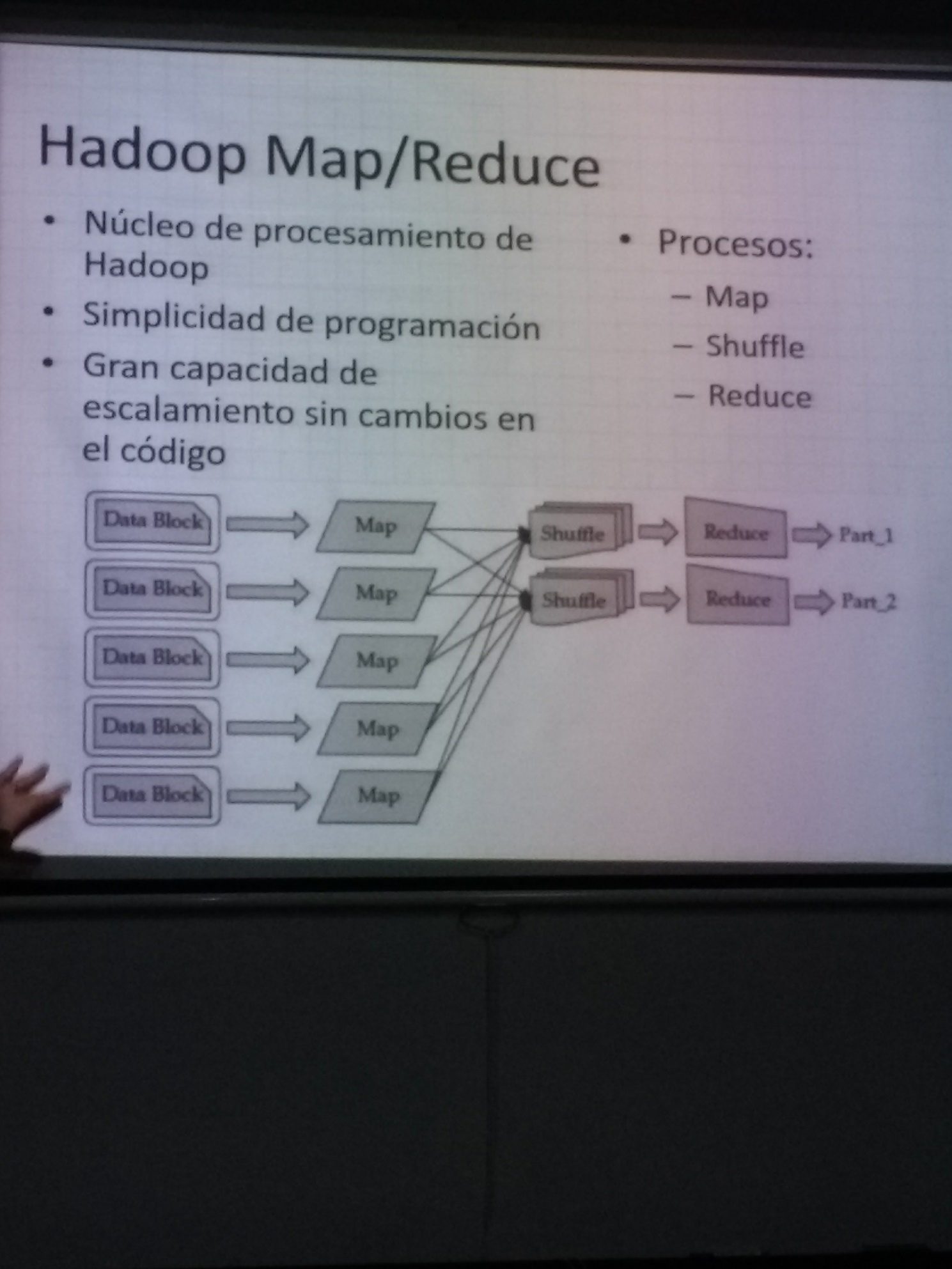
Hadoop Distributed File System

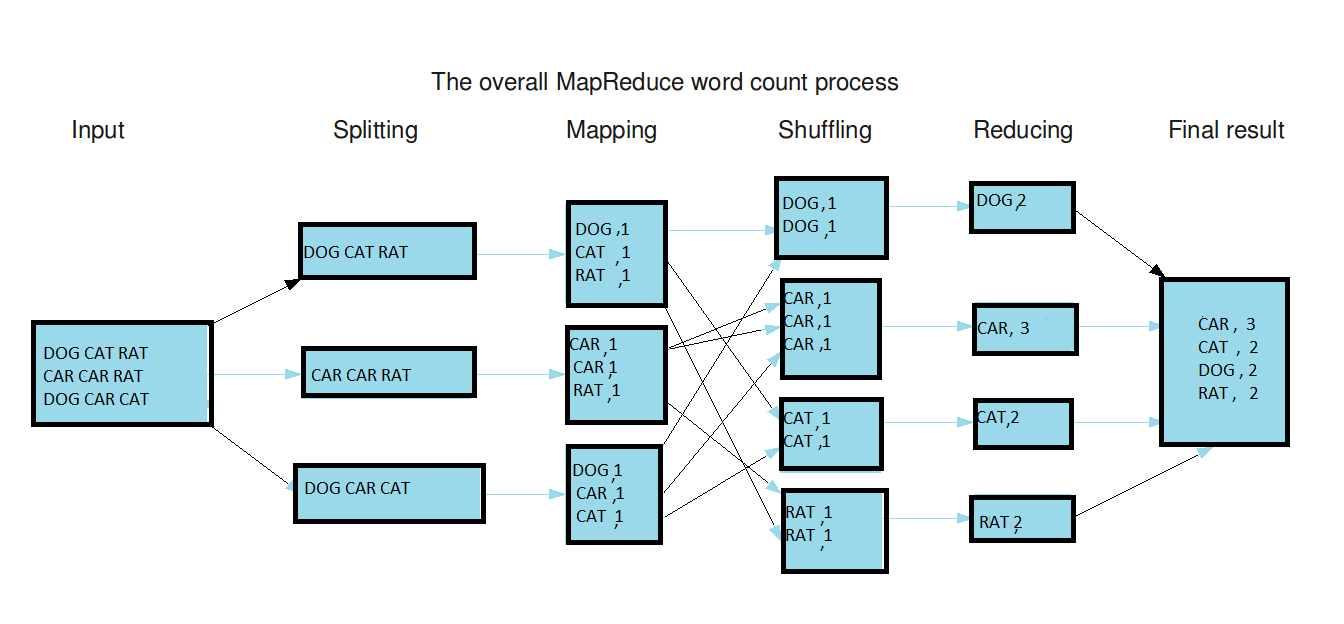
* Sistema de archivo distribuido, tolerante a fallos y basado en hardware de bajo porte (miles de PCs).
* Archivos se almacenan en clúster, en forma redundante (en PCs distribuidas por todo el mundo).
* Rapidez y facilidad de acceso.
* Datos se dividen en bloques que se almacenan en clustered, permitiendo procesamiento paralelo (cada PC puede intervenir en el procesamiento).
* *Componentes:*
  + **Name Node:** nodo que administra la metadata y controla medianamente el sistema de modo tal que sea auto-recuperable.
  + **Data Node:** conjunto de paquetes. Un conjunto de Data Nodes conforman un Rack.
  + **Block:** unidad mínima en que se divide el paquete de datos a procesar.



Hadoop Map/Reduce

* Núcleo de procesamiento de Haddop.
* Simplicidad de programación.
* Gran capacidad de escalamiento sin cambios en el código.
* Procesos: cada uno de los pasos se almacena en el mismo sistema de archivos.
  + Map: toma un bloque y lo procesa. Se detecta si es útil o no, y se le otorga un único sentido de clasificación, único criterio. Ejemplo: una imagen, la procesa y la mapea con el criterio “caribe.
  + Shuffle: procesos del framework que unifican todos los paquetes mapeados para cada criterio.
  + Reduce: procesa los paquetes unificados para obtener reportes y estadísticas.





Ejemplo de aplicación

Problema de monitoreo de infraestructura aplicativa.