UTN-FRBA

GESTION DE DATOS

DATA WAREHOUSE Y DATA MINING

DIRECTOR CATEDRA: ING. ENRIQUE REINOSA

INTRODUCCION

- La inteligencia del negocio consiste en la Transformación de datos en información y, ésta, en conocimiento, con la intención de mejorar al máximo el proceso de toma de decisiones de la organización.

- Desde la perspectiva de la tecnología de la información, la inteligencia del negocio se define como "el conjunto de metodologías, herramientas y estructuras de almacenamiento" que permiten la reunión, depuración y transformación de los datos en una información integrada que se pueda analizar y convertir en conocimiento para la optimización del proceso de toma de decisiones.

SISTEMAS DE DATA WAREHOUSE

- A partir de la década del ‘90, los sistemas de Data Warehouse se convirtieron en el centro de la arquitectura de los Sistemas de Información y surgieron para resolver los problemas que acarreaba **la extracción de información sintética desde datos atómicos almacenados en bases de datos de producción.**
- Un sistema de Data Warehouse incluye las siguientes funcionalidades:
  - **Integración de bases de datos heterogéneas (relacionales, documentales, geográficas, archivos, etc.).**
  - **Ejecución de consultas complejas no predefinidas que visualicen el resultado en forma de gráfica y en diferentes niveles de agrupamiento y totalización de datos.**
  - **Agrupamiento y desagrupamiento de datos en forma interactiva.**
  - **Análisis de problemas en términos de dimensiones. Por ejemplo, permite analizar datos históricos a través de una dimensión tiempo.**
  - **Control de calidad de datos para asegura la consistencia de la base y la relevancia de los datos que contribuirán con la toma de decisiones**

CONCEPTO DE DATA WAREHOUSE

- ***Data Warehouse (DW)***: es una base de datos corporativa cuya característica principal es la integración y el filtrado de información de una o varias fuentes, que luego procesará para su análisis desde diferentes punto de vista y con una gran velocidad de respuesta. Es una colección de datos históricos e integrados diseñada para soportar el procesamiento informático para la tomas de decisiones estratégicas que no utilizan para la operatoria diaria

CARACTERISTICAS DE UN DW

- **Está orientado a sujetos:** no se orienta a los procesos u operaciones clásicas, como en el caso de los sistemas y diseños transaccionales. Su modelo operacional orientado a los sujetos mayores de la organización se diseña alrededor de operaciones y funciones.
- **Es integrado:** Esto significa que los datos, cuando se mueven desde el ambiente transaccional u operacional, se integran antes de ingresar en DW.
- **Es temático:** Desde el entorno operacional, solamente se añadirán los datos que se necesitan en el proceso de generación de conocimiento del negocio. Estos datos, distribuidos por temas para facilitar la comprensión de los usuarios finales, se pueden reunir en una tabla de DW. Como toda la información se encuentra en un mismo lugar, los requerimientos de información acerca de los clientes se responderán sin complicaciones.

DUDA (??)

Se integran entre si TODOS los datos antes de ingresar al DW?

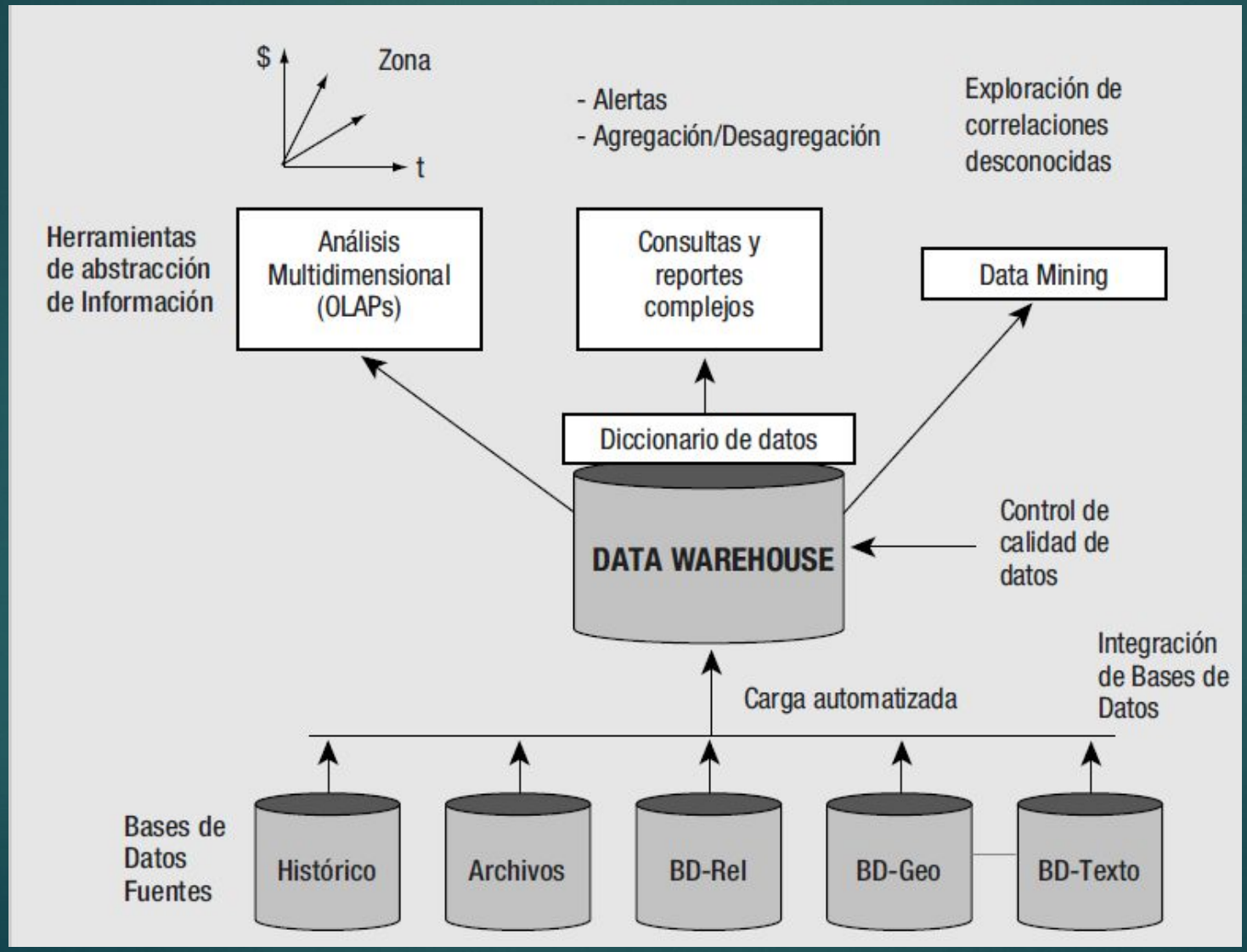
CARACTERISTICAS DE UN DW

- **Es variante en el tiempo:** Los datos en DW varían en el tiempo. Esto significa que son rigurosos en un determinado momento y no en otro. La variación de los datos se expresa de diversas maneras a través de un largo horizonte temporal.
- **Es simple de manejar:** En una base de datos transaccional, los *updates*, *inserts* y *deletes* se afectan a los datos operacionales. DW, por el contrario, opera los datos de una forma más simple, ya que sólo necesita dos operaciones: la carga inicial y el acceso a los datos. En este caso no se necesitan *updates*.
- **No es volátil:** El almacén de información de un DW se puede leer pero no admite ninguna modificación. En consecuencia, la información es inalterable y sus actualizaciones no la cambian. Sólo se incorporan las últimas variables.

DUDAS (???)

1. Que serian los "datos operacionales"?
2. Menciona que afecta a los datos operacionales, porque son db relacionales?

ARQUITECTURA DE UN DW



OBJETIVOS DE UN DW

- El objetivo de un ambiente de Data Warehouse consiste, principalmente, en la conversión de los datos de las aplicaciones del ambiente transaccional (OLTP) en datos integrados de gran calidad. Luego, es necesario que se los almacene en una estructura que facilite el acceso de los usuarios finales en un ambiente destinado a la toma de decisiones (OLAP).
- Durante este proceso, la totalidad de los datos se resumen y se incorporan aL DW, es decir, se los transfiere de manera periódica, de acuerdo con el análisis de negocios que se esté tratando.

Calidad lo menciona en la pag. 11 y 21

FUNCIONALIDADES DE UN DW

- Las funcionalidades de DW se pueden subclasificar en cinco grandes grupos: cada uno de ellos es responsable de un conjunto de procesos específicos, indispensables para el ambiente de soporte destinado a la toma de decisiones
  - Acceso a Fuentes (*Source*)
  - Carga (*Load*).
  - Almacenamiento (*Storage*).
  - Consultas (*Query*).
  - Utilización de Metadatos (*Meta Data*).

ACCESO A FUENTES

- Esta funcionalidad incluye los procesos que se aplican en las bases de datos fuentes a los datos que se transferirán. Si bien las bases de datos fuentes son las bases de datos operacionales de la organización, en la actualidad se las incluye, cada vez más, a las bases de distribución pública sobre industria, demografía y clientes potenciales. Estos datos llegan, muchas veces, de diferentes fuentes.
- Entre un 73 y 80% del tiempo de desarrollo de DW se destina durante la fase de análisis y diseño a los procesos asociados con la función de acceso a fuentes como, por ejemplo, **mapeo**, **integración** y **muestreo** de datos.
- Los factores que impactan directamente sobre el tiempo destinado a estas actividades son: el número de aplicativos fuentes que serán mapeados a Data Warehouse, la calidad de los metadatos mantenidos en esas aplicaciones y las reglas de organización que las gobiernan.).

CARGA

- La funcionalidad abarca diferentes procesos:

  - ***Extracción:*** es el primer paso de la preparación de los datos y comprende el acceso a los datos de los aplicativos. Para la extracción existen diferentes alternativas que equilibran la *performance* y las restricciones de tiempo y de almacenamient.
  - ***Depuración:*** es el proceso que verifica la calidad de los datos.
  - ***Conversión:*** es el último paso en la preparación de los datos que se cargarán en el DW. Este proceso necesita reglas de conversión de valores de aplicativos locales a globales e integrados.
  - ***Carga de datos*** es el proceso que ingresa los datos al DW

DUDA (???)
Como verifica la calidad?

ALMACENAMIENTO

Manejadores de DB

DBMS, -> Data base management system
RDBMS -> Relational data base management system
MDBMS -> Multidimensional DataBase Management System

- El almacenamiento abarca la arquitectura que se necesita para incluir varias vistas en DW. Si bien se suele decir que Warehouse es un único almacén, en realidad, sus datos pueden estar desperdigados en muchas bases que se manejan a través de diferentes DBMS's.
- Los manejadores que se ajustan a esta tarea son dos: los relacionales (RDBMS's) y los multidimensionales (MDDBMS's).
- En el caso de los MDDBMS, los datos se organizan en un *array* de *n* dimensiones. Cada una de ellas representa un aspecto del negocio que se analizará.
- Muchas veces, las diferentes áreas de una organización necesitan sistematizar sus respectivas visiones de los negocios como un *array* multidimensional que optimice sus requerimientos específicos. Sin embargo, no se aconseja que los requerimientos de todas las áreas sean soportados por la misma base multidimensional.

DUDA (???)

Que se aconseja entonces? Que cada area tenga su MDB?

CONSULTAS

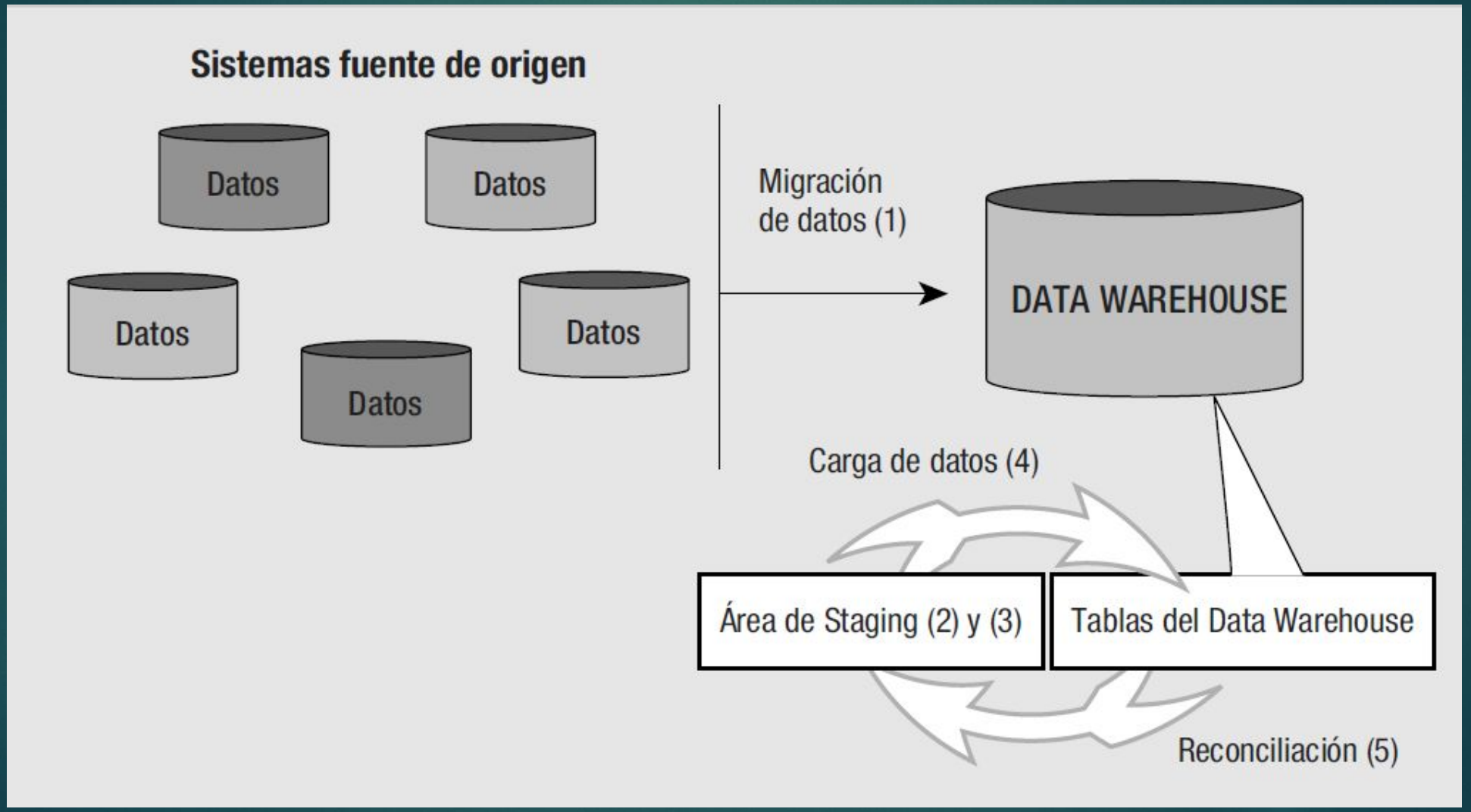
- El ambiente de consultas mediante sus herramientas OLAP permite que el usuario dirija el análisis y la producción de reportes. Nuevas tecnologías prometen soportar la nueva generación de herramientas de análisis:

  - **Data Mining:** se encargan del análisis de los datos para verificar la existencia de correlaciones inesperadas entre ellos.
  - **Simulación de negocios:** crean las herramientas necesarias para comprobar el impacto de las transformaciones en el ambiente negocios y establecen, si se considera conveniente, nuevas reglas de organización que realimentarán los aplicativos operacionales.

METADATOS

- Un **metadato** es "toda aquella información descriptiva sobre el contexto, calidad, condición o características de un recurso, dato u objeto **que** tiene la finalidad de facilitar su recuperación, autentificación, evaluación, preservación y/o interoperabilidad"
- El conocimiento de los Metadatos es tan esencial como el conocimiento de los datos de Data Warehouse. Incluyen el dominio, las reglas de validación, la derivación y la conversión de los datos extraídos.
- Los metadatos deben estar disponibles para el análisis que realizan los usuarios. En este caso, los administradores pueden manejar y proveer el acceso a través de los servicios de repositorio

MIGRACION DE DATOS



MIGRACION DE DATOS

- La ***migración*** es trasladar los datos desde los sistemas seleccionados de origen hasta el *stage* de DW. Sólo se moverán los datos solicitados por los usuarios para la emisión de reportes o aquellos que se utilizan durante los procesos de conversión y carga, de esta manera se previene el ingreso de información innecesaria.
- Los datos que se moverán al *stage* de DW incluirán datos referenciales y transaccionales. Por ejemplo, en un DW de ventas, los datos referenciales se relacionarán con la información del cliente y los transaccionales serán la información asociada con la venta a un cliente.
- Lo más importante es entender dónde se ubicarán los datos y cuáles se reubicarán. No se debe subestimar esta tarea; en caso de duda, es mejor que se deje los datos afuera, a menos que se sepa qué se hará con ellos.

DEPURACION DE DATOS

- La ***depuración*** de datos es corregir para estandarizar el formato y completar cualquier valor requerido por DW. Este proceso contribuye con la identificación de los datos redundantes que, durante el proceso de carga, no se ingresarán en DW. Para ello, se utilizan herramientas de *software* que migren, depuren y conviertan los datos. El retorno de la inversión justifica la compra de herramientas de *software* en vez de desarrollar *scripts* en SQL.
- Los costos asociados a mantener y ampliar desarrollos propios de *scripts* SQL excederán significativamente al de comprar herramientas de *software* desarrolladas por terceros.

DEPURACION DE DATOS

Tabla 5.1:

Nombre	Apellido	Empresa	Area	Telefono	Provincia
ANDRES	PEREZ	IBM	11	5355-2299	BA
Andres	Perez	I.B.M,	11	5355-2299	BA
Andrés	Pérez	IBM SA	11	5355-2299	
Andes	Peres	I.B.M. S.A.	11	5355-2299	BA
Martín	López	Bunge S.A.	351	272-2700	CO

Tabla 5.2:

Nombre	Apellido	Empresa	Area	Telefono	Provincia
Andrés	Pérez	I.B.M. S.A.	11	5355-2299	BA
Andrés	Pérez	I.B.M. S.A.	11	5355-2299	BA
Andrés	Pérez	I.B.M. S.A.	11	5355-2299	BA
Andrés	Pérez	I.B.M. S.A.	11	5355-2299	BA
Martín	López	Bunge S.A.	351	272-2700	CO

CONVERSION DE DATOS

proceso previo a la
a cargarlos al DW

- El objetivo de la ***conversión*** de los datos es cambiar los datos con el formato y la estructura requeridos por el DW. El proceso de conversión debería reducir el número de elementos de datos que se cargan desde el *stage* del DW.
- En el desarrollo de las reglas de conversión para este proceso, sólo se utilizarán aquellos elementos de datos que se requieran para DW. Si existieran otros que resultaran innecesarios, se prevendrá su ingreso en DW y no se los incorporará en las sentencias de conversión o carga.

CARGA DE DATOS

proceso previo a la
a cargarlos al DW

- La ***renovación completa*** comienza truncando las tablas en Data Warehouse y luego cargándolas con todos los datos requeridos. Esta alternativa puede prevenir que datos no deseados ingresen a Data Warehouse abarcando condiciones en las sentencias de carga

- La ***renovación incremental*** identifica los cambios que se produjeron en los datos origen desde la última vez que se cargó Data Warehouse y, luego, inserta, actualiza o borra registros de datos en cada tabla de Data Warehouse como se lo solicite.

CONCILIACION DE DATOS

- El proceso de conciliación identifica los problemas de datos que, si no se les diera importancia, pasarían los controles de prevención. Este proceso se diseña para proveer veracidad y para la identificación de los datos que no concuerdan con la información que contiene el sistema de origen. La conciliación de los datos determina la precisión y la integridad de la información. Para ello se debe analizar:
  - **La calidad de datos:** la exactitud se evalúa con el uso de totales de control sobre los elementos de datos seleccionados, que luego se compararán con los resultados anticipados.
  - **La cantidad de datos:** la integridad se determina cuantificando el número de registros y comparando los resultados con el número de registros anticipados.
- Independientemente del enfoque que se utilice, la conciliación de DW proveerá una red segura que identificará las excepciones de datos y asistirá con preguntas y perspectiva de direccionamiento en todas las partes interesadas dentro de la organización.

CONCILIACION DE DATOS

- ***Conciliación completa*:** Al finalizar cada proceso de carga, se realiza una conciliación completa que compara la información de Data Warehouse con la del sistema origen correspondiente.
- ***Conciliación por Fase*:** La conciliación se realiza después de cada etapa del flujo del proceso de datos, cuando no es factible una conciliación completa porque debido al número de sistemas de origen o a la complejidad de los procesos de depuración o conversión.
- Con la conciliación por fase, se determinan la veracidad e integridad de los datos luego de cada una de las siguientes etapas:

CONCILIACION POR FASE

- **Migración de datos:** después de que los datos del sistema origen han sido migrados al *stage* deL DW, se realiza la conciliación entre los datos del sistema origen y los del *stage* deL DW.
- **Depuración:** cuando termina el proceso de depuración, se realiza la conciliación entre los datos no depurados, el listado de excepciones y los datos depurados del *stage* del DW.
- **Conversión:** una vez que finaliza el proceso de conversión, se produce la conciliación entre los datos depurados, la lista de excepciones y los datos convertidos del *stage* del DW.
- **Carga:** después de terminado el proceso de carga, se hace la conciliación entre los datos convertidos del *stage* del DW.

DATA MARTS

- Se denomina Data Warehouse a un almacén de datos integrado; Data Marts, a las vistas multidimensionales de cada área.
- Los **Data Marts** se ajustan mejor a las necesidades que tiene una parte específica de un negocio, más que a las de toda una organización. Optimizan la distribución de información útil para la toma de decisiones y se enfocan al manejo de datos resumidos o de muestras, más que a la historia presentada con detalle.
- Los Datas Marts deben su popularidad a que disminuyen de manera significativa los costos asociados a su creación y operación.

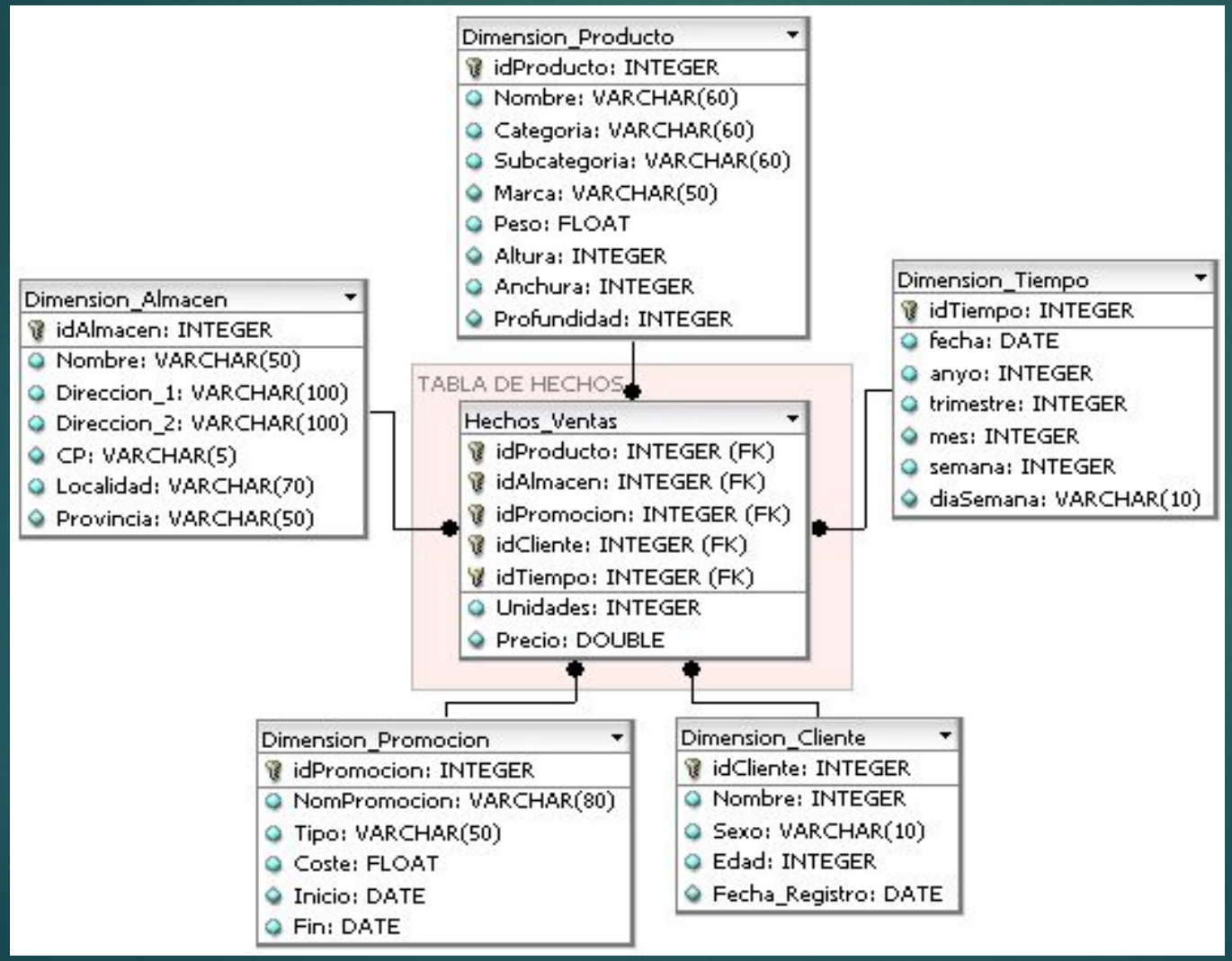
IMPLEMENTACION EN UN RDBMS

- Un DW o un Data Marts puede ser implementado en DBMS's Multidimensionales o Relacionales.

- Para implementar un DW en un RDBMS se utiliza lo que se denomina el ***Modelo Estrella (STAR MODEL)***

- ***Modelo Estrella:*** es un modelo de datos conformados por dos tipos de tablas, los hechos y las dimensiones.

MODELO STAR


Dimension_Producto
idProducto: INTEGER
Nombre: VARCHAR(60)
Categoria: VARCHAR(60)
Subcategoria: VARCHAR(60)
Marca: VARCHAR(50)
Peso: FLOAT
Altura: INTEGER
Anchura: INTEGER
Profundidad: INTEGER
Dimension_Almacen
idAlmacen: INTEGER
Nombre: VARCHAR(50)
Direccion_1: VARCHAR(100)
Direccion_2: VARCHAR(100)
CP: VARCHAR(5)
Localidad: VARCHAR(70)
Provincia: VARCHAR(50)
TABLA DE HECHOS
Hechos_Ventas
idProducto: INTEGER (FK)
idAlmacen: INTEGER (FK)
idPromocion: INTEGER (FK)
idCliente: INTEGER (FK)
idTiempo: INTEGER (FK)
Unidades: INTEGER
Precio: DOUBLE
Dimension_Tiempo
idTiempo: INTEGER
fecha: DATE
anyo: INTEGER
trimestre: INTEGER
mes: INTEGER
semana: INTEGER
diaSemana: VARCHAR(10)
Dimension_Promocion
idPromocion: INTEGER
NomPromocion: VARCHAR(80)
Tipo: VARCHAR(50)
Coste: FLOAT
Inicio: DATE
Fin: DATE
Dimension_Cliente
idCliente: INTEGER
Nombre: INTEGER
Sexo: VARCHAR(10)
Edad: INTEGER
Fecha_Registro: DATE

IMPLEMENTACION EN UN RDBMS

- ***Tabla de Hechos (Fact Table):*** registra medidas o métricas de un Evento específico, generalmente consisten de valores numéricos (datos asociados específicamente con el evento), y claves foráneas que referencian a tablas de datos dimensionales que guardan información descriptiva. Se diseñan para contener detalles uniformes a bajo nivel (referidos como "granularidad" o "grano"), o sea que los hechos pueden registrar eventos a un gran nivel de atomicidad-
- ***Tabla de Dimensiones (Dimension Table):*** Las Dimensiones pueden definir una amplia variedad de características. Las tablas de Dimensiones generalmente tienen un bajo número de registros, en comparación a las tablas de hechos, pero cada registro puede tener un gran número de atributos para describir los datos del hecho.

DATA MINING

- Data Mining es un conjunto de técnicas que se utilizan para la obtención de la información implícita en grandes bases de datos.
- Data Mining se encarga, a través de un conjunto de herramientas y técnicas algorítmicas, de buscar los patrones de interés ocultos, que son los que permiten la anticipación de futuros acontecimientos gracias a la predicción de acontecimientos o al pronóstico de situación con cierto grado de probabilidad
- Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque está fuera de sus expectativas.

CARACTERISTICAS DE DM

- Los procesos de Data Mining corren sobre bases de datos de gran volumen; esto se produce por dos aspectos fundamentales que se analizarán a continuación:
  - **Gran cantidad de columnas:** cuantas más columnas se especifiquen en DW, mayor será el nivel de análisis y de detalle en Data Mining, dado que realiza diferentes combinaciones entre los patrones especificados —en este caso, las columnas predefinidas—. Entonces, la cantidad de conclusiones que entregue estará en estrecha relación con el nivel de combinación que realice.
  - **Gran cantidad de filas**: para que Data Mining pueda contrastar los resultados con más tiempo para, de esta manera, disminuya la cantidad de errores de estimación y desvíos, se necesita que las tablas tengan la mayor cantidad de filas posibles que provean toda la información histórica disponible.

CARACTERISTICAS DE DM

- Para que Data Mining se pueda ejecutar y cumplir con su objetivo, debe tener las siguientes características:
  - **Recolección de datos en gran escala:** unifica el contenido de la información de todas las bases de datos disponibles, internas o externas. Como ya se mencionó, se disminuyen los errores y desvíos si la información disponible contiene amplitud y profundidad porque, de este modo, mayor será la aproximación o proyección que se obtenga de la tecnología.
  - **Alta Tecnología y gran almacenamiento**: como Data Mining procesa un volumen de información considerable y realiza un importante número de combinaciones, necesita múltiples y veloces procesadores; también requiere una gran capacidad de memoria RAM y secundaria debido a los procesos intermedios de recolección y combinación de datos e información que ejecuta esta tecnología.
  - **Algoritmos de Data Mining:** DM funciona con la aplicación de diversas herramientas algorítmicas que son las que permiten la búsqueda de información oculta.

CARACTERISTICAS DE DM

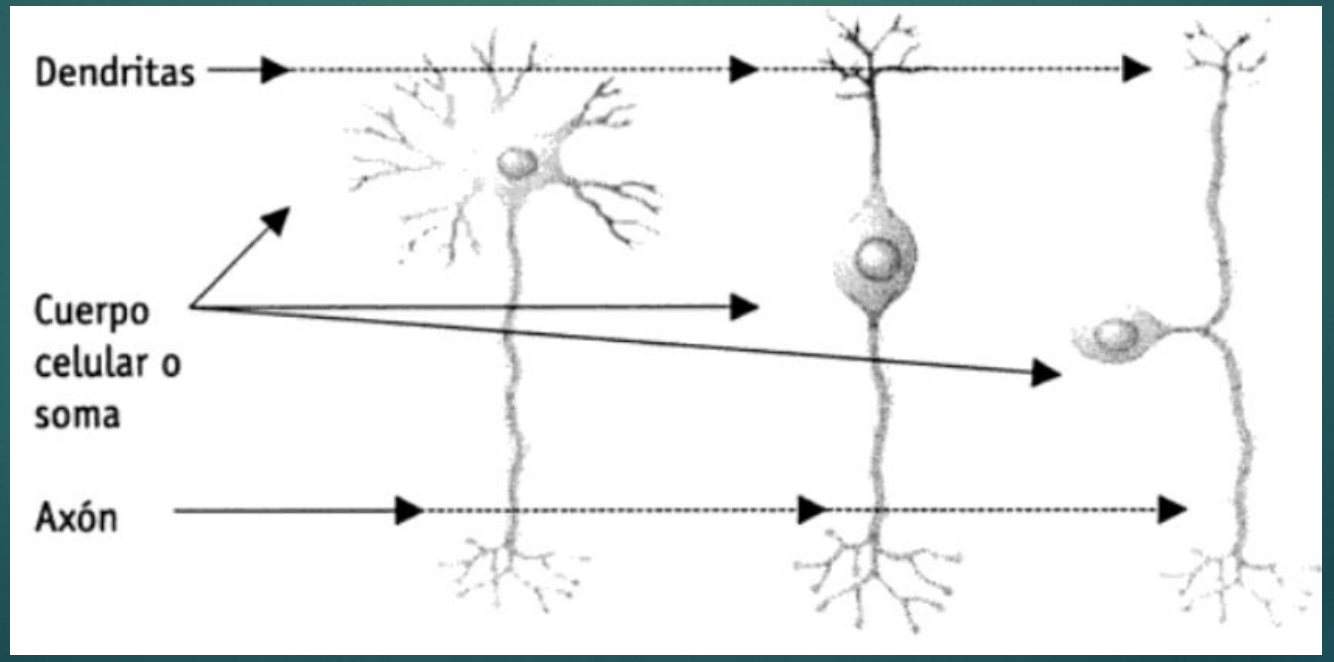
- La tecnología de Data Mining, con bases de datos de suficiente tamaño y calidad, genera nuevas oportunidades de negocios que proveen las siguientes capacidades:
  - **Predicción automatizada de tendencias y comportamientos**. Data Mining, al automatizar la búsqueda de información predecible en grandes bases de datos, puede inferir, ante una nueva situación o estímulo determinado, cuál sería el comportamiento futuro
  - **Obtención automatizada de modelos previamente desconocidos.** Para la identificación de los nuevos patrones de tendencia, es necesario la utilización de DM para que barra de un solo paso, a través de sus herramientas algorítmicas de búsqueda de información oculta, el DW e identifique los patrones desconocidos por la organización. Para descubrirlo evalúa **todos** los parámetros que conforman el comportamiento de los actores, y los combina con el fin de obtener una heurística de comportamiento que identifique, con una determinada probabilidad, el interés de los clientes por ese producto.

HERRAMIENTAS ALGORITMICAS

- Dentro de estas técnicas, las más utilizadas son:
  - **Redes neuronales artificiales:** son modelos predecibles, de características no lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.
  - **Algoritmos genéticos:** son técnicas de optimización con un diseño basado en el concepto de evolución y que utilizan procesos como las combinaciones genéticas, las mutaciones y la selección natural.
  - **Arboles de decisión:** estructuras cuya forma representa la copa de un árbol y que representan conjuntos de decisiones. Estas decisiones son las que generan las reglas que clasifican un conjunto de datos, que se segmentan mediante búsquedas arboladas. Dentro de los métodos específicos de árboles de decisión se incluyen, también, los Arboles de Clasificación y Regresión.

REDES NEURONALES

- La red neuronal artificial es un método de resolución de problemas que, como indica su nombre, emula el modo de conexión de las neuronas del cerebro. Esta red posee capas de unidades procesadoras —nodos— que se unen por conexiones direccionales.

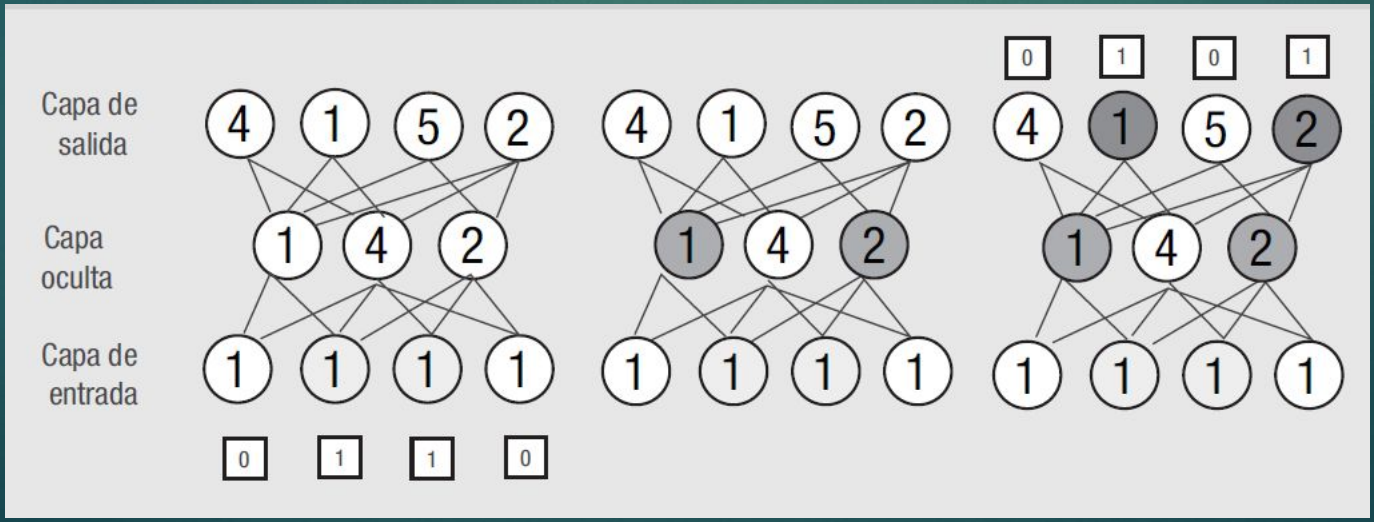


REDES NEURONALES

- Si la suma de todas las entradas que ingresan en una de estas neuronas virtuales es mayor que el famoso umbral de activación de la neurona, ésta se activa y transmite su propia señal a las de la siguiente capa.
- Por lo tanto, el patrón de activación se propaga hasta que llega a la capa de salida que lo devuelve como solución a la entrada presentada. De la misma manera que en el sistema nervioso de los organismos biológicos, con el transcurso del tiempo, una red neuronal aprende y afina su rendimiento gracias a la repetición de rondas en las que ajusta sus umbrales hasta que, para cualquier entrada, la salida real coincide con la deseada.
- Este proceso, denominado entrenamiento de la red, lo puede supervisar un experimentador humano, o puede correr automáticamente con un algoritmo de aprendizaje que los optimice de manera constante.

REDES NEURONALES

- Aquí vemos una sencilla red neuronal con tres capas: la de entrada, con cuatro neuronas; la oculta, con tres y, por último, la de salida, con cuatro. El umbral de activación de cada neurona se representa por su número. Para que se excite debe recibir, por lo menos, esa cantidad de entradas. El diagrama muestra cómo la red neuronal recibe una cadena de entrada y cómo la activación se extiende por la red hasta producir una salida.



REDES NEURONALES

- Existen algoritmos de optimización:
  - **Ascenso a colina** o **voraces:** Son similares a los genéticos pero con una mayor sistematización y menor aleatoriedad.
  - **Recocido simulado**: El recorrido simulado es parecida a los algoritmos evolutivos. Su nombre proviene del proceso industrial que consiste en calentar un material por encima de su punto de fusión y, luego, se enfría para eliminar los defectos en su estructura cristalina, que produce un entramado de átomos estable y regular

ASCENSO A COLINA

- Un algoritmo de ascenso a colina comienza con la solución de un problema que tiene a mano que, normalmente, se elige al azar. Después, la cadena se muta y, si ésta proporciona una solución con mayor amplitud que la anterior, se conserva la nueva; en caso contrario, la actual.
- Este algoritmo se repite hasta que no se pueda encontrar una mutación que provoque un incremento en la aptitud de la solución actual, y ésta se devuelve como resultado.
- Esta técnica se denomina ascenso a colina porque, en general, se representa con un paisaje en el que se encuentran todas las soluciones posibles de un problema particular. Cada solución, a la vez, se constituye por un conjunto de coordenadas de ese paisaje. La mejores soluciones están a mayor altitud y forman colinas y picos; las peores, a menor altitud y forman valles. Un "trepacolinas" es, en consecuencia, un algoritmo que se inicia en un punto de paisaje y se mueve hacia arriba de la colina.
- Este tipo de algoritmo también se lo denomina "voraz" porque siempre hace la mejor elección en cada paso, con la esperanza de que se obtendrá el mejor resultado global.

RECOCIDO SIMULADO

- En el recocido simulado una función de aptitud define una solución candidata. Esta clase de recorrido añade, además, el concepto de "temperatura", que es una cantidad numérica global que disminuye de manera gradual. En cada uno de sus pasos, esta solución muta.
- La aptitud de la nueva solución se compara con la anterior y, si es mayor, se la conserva. Si ocurre lo contrario, el algoritmo decide si la conserva o la descarta sobre la base de la temperatura. Si ésta es alta, se conservan incluso los cambios que causan disminuciones significativas en la aptitud y se utilizan para la siguiente ronda. Sin embargo, a medida que disminuye la temperatura, el algoritmo tiende a aceptar sólo los cambios que aumentan la aptitud.
- Finalmente, cuando la temperatura alcanza el cero y el sistema se "congela", la configuración que exista en ese punto se convierte en la solución.

ALGORITMOS GENETICOS

- Son algoritmos de optimización, o sea, tratan de encontrar la mejor solución a un problema dado entre un conjunto de soluciones posibles. Los mecanismos que utilizan los AG para llevar a cabo esa búsqueda consisten en procesos que se asemejan a la evolución biológica, de allí el nombre de algoritmos genéticos.
- Trabaja sobre el concepto de la ***mutación***, dada una población de soluciones, y en base al valor de la función objetivo para cada uno de los individuos (soluciones) de esa población, se seleccionan los mejores individuos (que son aquellos que minimizan la función objetivo) y se combinan para generar otros nuevos. Este proceso se repite cíclicamente hasta probar todas las combinaciones y encontrar la óptima.
- Es simillas con el proceso que se da en la naturaleza, en el que los individuos compiten por su supervivencia. Los mejor adaptados al medio, es decir, los que pueden optimizar la función objetivo, sobreviven y transmiten su material genético a las futuras generaciones

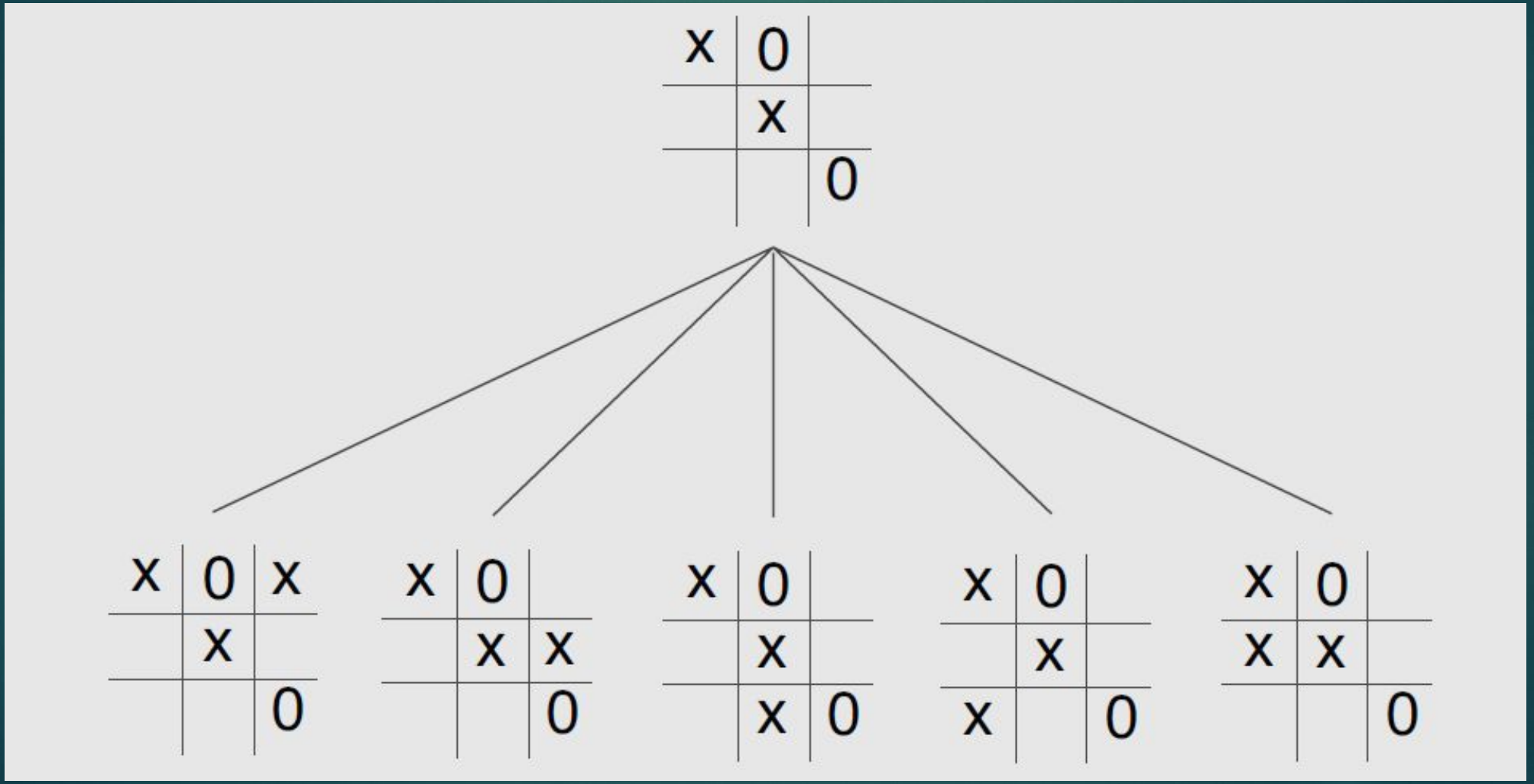
ALGORITMOS GENETICOS PASOS

- Definir la solución
- Filtrar
  - Aplicarle la función objetivo.
  - Ordenar los individuos en función de los valores obtenidos.
  - Seleccionar los mejores individuos (soluciones) para el cruce.
- Cruzar los individuos.
- Mutación de los descendientes.
- Inserción.
- Si se cumple la función objetivo ”terminar”, de lo contrario volver al paso 2.

ARBOLES DE DECISION

- Los árboles de decisión son una técnica de programación que permite analizar decisiones secuenciales basadas en el uso de resultados y probabilidades asociadas; es decir, en una heurística de ocurrencia.
- Los árboles de decisión, se utilizan en la Inteligencia Artificial, especialmente en los denominados **sistemas expertos**, que se basan en grandes bases de datos, en las que se cargan reglas de decisión que encuentran su fundamento en la experiencia de los expertos en una ciencia determinada sobre la que versará el sistema. De esta manera, se lo puede utilizar para establecer un diagnóstico determinado, en el que se evalúa todos los caminos posibles dentro del árbol.

ARBOLES DE DECISION



ARBOLES DE DECISION

- Las ventajas de un árbol de decisión son:
  - Resume los ejemplos de partida y permite la clasificación de nuevos casos siempre y cuando no existan modificaciones sustanciales en las condiciones que generaron los ejemplos que sirvieron para su construcción.
  - Facilita la interpretación de la decisión adoptada ya que permite regenerar el camino decisorio aplicado.
  - Proporciona un alto grado de comprensión del conocimiento utilizado en la toma de decisiones.
  - Explica el comportamiento respecto a una determinada tarea de decisión.
  - Reduce el número de variables independientes.
  - Es una magnífica herramienta para el control de la gestión empresarial.

VENTAJAS DEL DM

- Contribuye con la toma de decisiones estratégicas y proporciona un sentido automatizado para identificar información clave desde volúmenes de datos generados por procesos tradicionales y de *Business Intelligence*.
- Permite a los usuarios dar prioridad a decisiones y acciones e indica los factores que tienen una mayor incidencia, qué segmentos de clientes son desechables y qué unidades de negocio son sobrepasadas y por qué.
- Genera modelos descriptivos: en un contexto de objetivos definidos en los negocios, permite a las organizaciones, sin que se considere la industria o el tamaño, explorar automáticamente, visualizar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales.
- Genera modelos predictivos: permite que relaciones no descubiertas e identificadas a través del proceso de Data Mining se expresen como reglas de negocio o modelos predictivos.