**Las Organizaciones Industriales – Introducción**

Raíces organizacionales

* Prototipo de fábrica de Adam Smith, precursor de la consultoría de negocios y del principio de división del trabajo.
* Trabajadores especializados. Menores requerimientos de capacitación, fácil incorporación.
* Mayor velocidad y producción. Se ahorra el tiempo de pasar de una tarea a la otra.
* Facilitó la invención de máquinas.
* Promovía el mejoramiento de las prácticas de los trabajadores en tareas específicas.
* Cada trabajador solo realiza una parte del trabajo; no tenía la visión general de la tarea.

El Ferrocarril

* Impulsó el desarrollo económico. Alcance de mercados lejanos.
* Tecnología de la Administración de Negocios. Tecnología Blanda. Burocracia Moderna.
* Líneas de autoridad y dependencia.
* Reglas preestablecidas, programadas. Planes de contingencia.

Henry Ford

* Refinó la división del trabajo. Llevó el trabajo hasta el trabajador.
* División extrema complejidad en coordinación y combinación de trabajo y resultados.
* El sistema administrativo no soportaba la magnitud de las industrias.

Alfred Sloan

* Cambio anual de estilos de los autos obsolescencia programada.
* Crea la Producción en Serie.
* Aplicó la división del trabajo a la administración: unidades pequeñas, descentralizadas y controlables.
* Gestionó las operaciones mediante el control de estadísticas y parámetros financieros.

La Planificación (Robert McNamara - Ford, Harold Geneen – ITT, Regionald Jones - General Electric)

* Desarrollan las técnicas de planificación: determinación de los negocios, capital y utilidades esperadas.
* Surgen controladores, auditores y planificadores.
* Recaban información de inventario y actividades para ajustar la planificación.
* Técnicas para presupuestar, controlar y planificar.

Fin 2º Guerra Mundial

* Demanda insaciable de bienes y servicios.
* Para crecer, se agregaban trabajadores en la base del organigrama. Principal preocupación capacidad.

Nuestros días

* Procesos más complejos y con más tareas.
* Mayor nivel de personal en niveles intermedios del organigrama.
* Mayor distancia entre la administración y los clientes.
* Avance de la tecnología. Desaparición de las fronteras.

Nuevos Tiempos: 3 fuerzas básicas que impulsan a producir un cambio:

* **Clientes**: los clientes asumen el mando en la relación vendedor-cliente.
  + Antes: los clientes se adaptaban. Poca oferta. Hoy: la empresa se adapta al cliente. La oferta es mucha.
  + Se exige personalización. Mayor diversificación.
* **Competencia**: la competencia se intensifica, aparecen nuevos competidores.
  + No existen territorios protegidos de competencia extranjera.
  + Un competidor eficiente sube el umbral competitivo para los demás.
* **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 para innovar.

Conclusión

* Separación entre actividades relacionadas a la producción y a la gestión.
* Categorías de procesos: completamente estructurados, completamente no estructurados.
* Volumen de información empresarial en crecimiento constante. Desafío: encontrar información valiosa oculta.
* El éxito de las organizaciones depende de la innovación continua, ofrecer cosas nuevas.
* La tecnología es una posibilitadora de la administración de la información de los procesos.

**Business Intelligence (BI)**

* Estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización.
* Objetivo: establecer *estrategias, fortalezas* y ­*debilidades.*
* Evolución de los sistemas de información tradicionales operacionales. No sustitutiva, complementaria.

Conceptos

* **Dato**: representación simbólica de un atributo. No tiene un valor semántico. *Ejemplo*: sueldo de un docente.
* **Información**: datos procesados. Tiene significado, relevancia y propósito. Los datos se convierten en información cuando se les añade significado. *Ejemplo*: es susceptible de adquirir un préstamo X a pagar en 12 cuotas.
* **Conocimiento**: mezcla de experiencia, información y “saber hacer”. Útil para la acción. *Ejemplo*: quienes tienen hijos, casa y superan los procesos administrativos, son buenos candidatos para la entrega de préstamos.

Sistemas Operacionales vs Business Intelligence

*Sistemas Operacionales*

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

*Business Intelligence*

* Optimizadas para *preguntar*.
* Datos desnormalizados para 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**. Herramientas de análisis para seleccionar y manipular datos de interés.
* **Orientación al usuario final**. Independencia de conocimientos técnicos para acceder a la información.

**Procesos de decisión**

Elegir entre varias alternativas con el fin de alcanzar objetivos. Resolver problemas + investigar oportunidades de negocio.

Taxonomía de Simon

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

* *Estructurado:* automatizable, fases estructuradas, procedimientos para obtener la mejor solución.
* *No estructurado:* no automatizable, sin fases estructuradas, se resuelven mediante la intuición humana.

Los semi-estructurados tienen alguna fase estructurada. Se resuelven combinando procedimientos y juicio humano.

Soporte Informático para Procesos de Decisión Estructurados

Aplicados a procesos repetitivos, automatizables, tipificables.

Soporte Informático para Procesos de Decisión No Estructurados

Sistemas de información interactivos que permiten utilizar datos y modelos para resolver problemas. No resuelve problemas por sí solo. Propósito: ayudar al decisor, no reemplazarlo. Ejemplos: análisis multidimensional, data mining, forecasting, tablero de comando, gestión de conocimiento.

**Conceptos Importantes**

**Data Warehouse**

Base de datos corporativa caracterizada por integrar y depurar información de una o más fuentes distintas, para luego procesarla y permitir su análisis desde diversas perspectivas, con alta velocidad de respuesta.

Características específicas

* Desnormalizada.
* Optimizada para las consultas. Pensado para responder rápidamente consultas complejas.
* No volátil. Es importante mantener toda la información histórica.

Características de su desarrollo

* Dificultoso, costo de tiempo. Proceso incremental.
* Identificación de fuentes. Ejemplo: determinar en qué fuente X dato está más actualizado, es más fiel, sin errores.
* Se debe involucrar mucho con los procesos de toma de decisiones.

**Data Mining**

Conjunto de métodos estadísticos que proporcionan información (correlaciones o patrones) cuando se dispone de muchos datos. *Datos + Estadísticas = Información*.

Proceso de analizar datos desde diferentes perspectivas con el objetivo de resumirlos en segmentos de información útiles.

Permite a los usuarios analizar datos desde diferentes dimensiones, categorizando y resumiendo relaciones identificadas.

Permite trabajar con grandes cantidades de observaciones.

Aplicaciones

* Comercio/Marketing: identificación de patrones de compra, búsqueda de asociaciones entre clientes y características demográficas, identificación de comportamiento de diferentes estratos o categorizaciones.
* Bancos: detección de patrones de uso fraudulento de tarjetas de crédito, identificación de clientes leales.
* Medicina: identificación de terapias médicas adecuadas para las enfermedades, estudio de factores de riesgo.

**Big Data**

Almacenamiento de grandes cantidades de datos, y la utilización de procedimientos para encontrar patrones repetitivos dentro de los mismos.

Diferencias respecto a Data Warehouse

* Mayor cantidad de datos.
* Datos estructurados y no estructurados. Datos más complejos (texto, sonido).

Son complementarios.

Las tres “V” (dimensiones) de Big Data:

* *Velocidad:* para procesar y almacenar nueva información constantemente.
* *Volumen:* grandes cantidades de datos, crecimiento exponencial.
* *Variedad*.

Orígenes de los datos

Sistemas operacionales, transacciones, e-marketing y web, lectores biométricos y sensores, machine to machine.

**Data Warehouse**

Características

1. *Orientación a temas*

* BD operacional: orientado a las aplicaciones, procesos, funciones. Ejemplo: Cuenta Corriente, Plazo Fijo.
* Data Warehouse: orientado a temas y procesos de decisión. Ejemplo: Ventas, Productos, Logística.

Organizado en torno a temas principales, excluyendo datos no usados para tomar decisiones.

1. *Integración:* Fuentes de Datos del DW

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

1. *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.

1. *No volátil*

* BD Operacional: datos actualizados regularmente registro a registro.
* Data Warehouse: el dato es cargado y accedido, pero nunca se modifica.
* Consecuencias: manejo de datos mucho más simple. Elimina el proceso de actualización de datos en línea: backup, recuperación, transacción, integridad, 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.

Ventajas

* Permite personalización y uso local de recursos.
* Menor cantidad de datos.
* Libertad para elegir software de análisis.
* Menor costo unitario de procesamiento y almacenamiento.

Carga del Data Mart

Depende de la arquitectura usada:

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

**Metadatos del Data Warehouse y del Data Mart**

Metadatos: directorio de los datos. Ayuda al analista a localizar contenidos, guía el mapeo de datos y a los algoritmos de agregación/sumarización.

Información contenida: fuente de los datos, descripción de las transformaciones recibidas, información descriptiva del DM o DW (tablas, atributos, relaciones, etc.), definición de los términos usados.

**Soporte requerido para Crear y administrar el DW/DM**

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

**Almacenamiento de datos externos**

* Se almacenan en el DW si son para toda la organización (evita redundancia, incompatibilidad, costos).
* Se guarda también el origen del dato (fuente, fecha, cantidad, descripción).

**La estructura de datos en el DW**

*Agregado*: información agregada, importante para usuarios específicos

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

Metadatos

*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.

Flujo de los datos y población de usuarios

Procesos de agregación

Procesos de Extracción, Transformación y Carga

Procesos de archivado

Enorme cantidad de usuarios

Gran cantidad de usuarios

Pocos usuarios

Muy pocos usuarios

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

**Tareas iniciales del diseño de un DW**

Pensar como un conjunto de hechos relacionados. No exhaustivo (solo se tendrán las transacciones de interés), no exclusivo (puede existir información repetida, debido a que la información es tomada de diferentes fuentes).

Construir una Matriz: DMs (filas) x Dimensiones (columnas).Listar de 10 a 30 DMs de fuente única y las dimensiones.

Consensuar la matriz con los interesados. Conformar las dimensiones de forma clara y única para todos los hechos.

**Hechos y dimensiones**

Los *hechos* contienen los datos de estudio. Las *dimensiones* contienen metadatos sobre los hechos.

Diseño de la tabla de Hechos

* Etapa 1: elegir un DM.
* Etapa 2: elegir el grano.
* Etapa 3: definir las dimensiones. Si una dimensión no coincide con el grano puede cambiarse el grano (recomendado) o descartar la dimensión.
* Etapa 4: definir los hechos.

**Modelos de diseño de un DW**

**Modelo de Inmon**

Modelo: DMs dependiendo del EDW (Enterprise Data Warehouse).

* EDW: contiene datos originales, sirve a las funciones de la empresa.
* DMs: desarrollados a partir del EDW, sirven a las funciones de los departamentos.

Características

EDW

* Datos muy detallados, de toda la organización.
* Historial: 1-10 años.
* Diseñado para uso futuro no conocido.
* Para toda la organización.
* Data-driven.

DMs

* Datos agregados, de interés del área.
* Historial de 60-90 días.
* Diseño para uso futuro concreto.
* Para uso del departamento.
* Requirement-driven.

Objetivos del diseño del EDW: mejorar la velocidad de acceso a los datos. Reflejar requerimientos generales.

Estructura de datos

Estructura de datos del EDW: modelo relacional ligeramente desnormalizado.

* Crear un lugar único para almacenar datos de uso frecuente.
* Crear redundancia de datos a usar regularmente junto a otros datos.
* Crear datos agregados cuando el dato es calculado y usado muchas veces.
* 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 departamentales.

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

**Modelo de Kimball**

* El DW es la unión de los DM.
* Usa modelo dimensional: mejor performance y comprensión.

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 mismo.
* **Monitoreo de cambios en los datos**: monitorea las fuentes e informa al ETL los cambios que considera de interés.

Mecanismos de monitoreo, dependiente del tipo de datos

* **Cooperación**: fuentes con triggers que permiten programar los cambios desde las mismas.
* **Registro de Operación**: log que puede ser inspeccionado para detectar cambios.
* **Consultas SQL**: para detectar cambios de interés.
* **Fotografías**: lecturas periódicas de datos. Por comparación de fotos sucesivas se detectan cambios. Más utilizado.

**Evolución del DW**

La arquitectura del DW debería permitir manejar cambios en las fuentes y en los requerimientos.

**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 |
| Autor | Inmon | Kimball |
| Énfasis | EDW | DMs |
| Diseño | Modelo normalizado basado en la empresa | Modelo dimensional DMs esquema de estrella |
| Data Set | EDW: datos atómicos. DMs: datos sumarizados | Contiene datos atómicos y sumarizados |
| Arquitectura | Niveles de áreas de interés y DMs dependientes | Área de interés y DMs |

**Modelado Dimensional**

Contiene la misma información que el modelo relacional. La diferencia radica en que el modelo dimensional almacena los datos de forma simétrica, con el objetivo de lograr: **comprensión** por parte del usuario, **performance** para consultas, **flexibilidad** para cambiar. Consta de tablas de dimensiones y de hechos.

Definiciones

* **Atributos:** generalmente describen una característica de una cosa tangible. Se organizan en **dimensiones**.
* **Hechos:** algo que se desea observar. Numéricos o textos.

Clasificación de hechos numéricos

* **Aditivos:** se pueden sumar en todas las dimensiones.
* **Semi-aditivos:** se pueden sumar en todas las dimensiones excepto en la dimensión tiempo, donde se suman y se divide por el número de períodos de tiempo.
* **No Aditivos:** medidas de intensidad no aditivas en todas las dimensiones.

Esquemas de modelos dimensionales

* **Esquema estrella**. Ventaja: consultas sencillas, utilizando pocos JOINS.
* **Esquema copo de nieve**. Las dimensiones se relacionan con otras dimensiones. Consultas más lentas/complejas.

Dimensiones Conformadas

* Dimensiones comunes a toda tabla de hechos que la utiliza.
* Grano: definir el nivel de registro individual de una tabla de hechos.

Clasificación de Hechos

* Básicos: representados por un campo de una tabla de hechos. Ejemplo: precio, cantidad en una venta.
* Derivados: combinación de hechos con alguna operación, que también residen en una tabla de hechos.
  + Aditivo: generado con hechos de una misma tabla. Comparte el mismo grano.
  + No aditivos: valor acumulado expresado en diferente grano. No puede mostrarse en la misma tabla.

Hecho Agregado

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

Se definen jerarquías de agregación (dimensiones) y niveles de agregación (campos dimensionales) para cada una.

Tablas Agregadas

* Por cada nivel de una jerarquía de agregación se puede generar una tabla de hechos que se denomina agregada.
* Se decide generarlas a partir de las estadísticas de uso de los datos por parte de los usuarios.
* Reducen el tiempo de respuesta.
* Se pueden eliminar cuando dejan de ser consultadas.

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

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

OLAP es sinónimo de vista multidimensional de datos.

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

*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.

Tecnología de Cubos

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

*Fortalezas:*

* Performance *óptima* para dar soporte a procesos de decisión. Acceso rápido a los datos.
* Ideal si se conoce un patrón de uso de los datos.
* Facilidad en la rotación del cubo. Permite examinar los datos desde muchas dimensiones.
* No es necesario escribir SQL.

*Debilidades:*

* Maneja reducida cantidad de datos.
* Gran tiempo de carga.
* Estructura de datos no flexible.
* El soporte para vínculos dinámicos (triggers) 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.
* Alta performance de las consultas.
* Alta capacidad de cálculo de información de negocio.
* Soporta un amplio rango de aplicaciones.

*Beneficios para responsables de sistema:*

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

Variantes de los soportes OLAP

* ROLAP (Relational On-Line Analytical Processing): genera cubos multidimensionales dinámicamente al momento de realizar las consultas. Almacena los datos en estructuras relacionales. Método más usado.
* MOLAP (Multidimentional On-Line Analytical Processing): almacena los datos físicamente en estructuras multidimensionales. Estructuras de almacenamiento particulares y técnicas de compactación de datos.
* HOLAP (Hybrid On-Line Analytical Processing): híbrido entre MOLAP y ROLAP. Los datos agregados y calculados se almacenan en estructuras multidimensionales y los de menor nivel de detalle, en estructuras relacionales.

**Soporte Inteligente para Negocios**

**Data Mining**

Es un proceso analítico diseñado para explorar grandes cantidades de datos en búsqueda de patrones consistentes y/o relaciones entre variables, y luego validar el hallazgo aplicando los patrones encontrados a nuevos conjuntos de datos.

Proceso de Data Mining

1º Paso - Selección

* Selección del conjunto de datos a utilizar. Población muy grande: técnicas estadísticas de selección de muestra.
* Generar dos conjuntos de registros representativos: entrenamiento y validación.

2º Paso - Preproceso

* Analizar y determinar datos erróneos, incorrectos, faltantes. Estrategias: descartar, omitir, inferir, promediar.

3º Paso - Transformación

* Los datos pueden requerir transformaciones por diferentes razones: diferencias de escalas entre los mismos, presencia de valores cualitativos que requieren ser cuantificados en escalas.

4º Paso - Data Mining

* Procesamiento de los datos mediante técnicas apropiada a efectos de encontrar patrones, en caso de existir.

5º Paso - Interpretación

* Interpretar los patrones a efectos de generar conocimiento. Validar los patrones con otro conjunto de datos.

Modelos Data Mining

* Modelos de verificación: toma una hipótesis definida por el usuario y realiza un test de validación.
* Modelos de descubrimiento: 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 incorrectos o faltantes.*
* *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, asociaciones entre consumidores y características geográficas.
* Banco: detectar uso fraudulento de tarjetas de crédito, clientes fieles.
* Seguros: identificar patrones de conducta de clientes de riesgo, comportamientos fraudulentos.

Funciones de Data Mining

* *Clasificación*: el sistema aprende a clasificar datos, formulando reglas.
* *Asociación*: se asocian atributos de una relación a otros. Ejemplo: clientes que compran A y B, compran C.
* *Clustering/Segmentación*: partición de una base de datos en grupos similares según algún criterio o métrica.
* *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.

Técnicas y herramientas de uso frecuente en Data Mining

* Visualización: permite el análisis exploratorio de los datos. Hace más visibles los patrones.
* Análisis de Clusters: define grupos de una población sobre cada uno de los cuales se pueden realizar nuevos análisis.
* Inducción:
  + Árboles de decisión: usados para clasificar objetos. Se construyen usando *datos de entrenamiento.*
  + Basados en reglas “si X entonces Y”.
    - Son simples e intuitivas, no estructuradas y poco rígidas. Difíciles de mantener.
    - Tipos: proposicionales, asociación.
* Redes Neuronales: modelo matemático que captura el patrón de comportamiento de un conjunto de datos. Permite realizar interpolaciones. No explica los resultados, por lo que es difícil interpretarlos.
* Técnicas Estadísticas. Se pueden usar en varias etapas de un proceso de DM: limpieza de los datos, análisis exploratorio de los datos, selección de datos.

Técnica de clasificación

Tarea de aprender una función que mapea cada conjunto de atributos X a una clase predefinida Y.

* X: conjunto de atributos. Discretos o continuos.
* Y: atributo objetivo, clase, grupo. Discreto.

Uso

* *Modelo descriptivo:* describe características de una clase.
* *Modelo predictivo:* predice la clase a la que pertenece un objeto desconocido.

Proceso para generar un modelo de clasificación

**Algoritmo de aprendizaje**: para generar un modelo que relacionaX con Y.

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

**Validación del modelo – métricas**

* *Exactitud*: .
* *Tasa de error*: .
* *Precisión*: (buenos pagadores de verdad / los que se clasificaron como buenos pagadores).
* *Recall*: (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, nodo interno (una sola entrada y dos o más salidas), nodo final (una sola entrada y no tiene salidas).
* Los nodos finales representan las clases.
* Los nodos raíz e internos contienen condiciones de test de atributos (CTA) que permiten dividir el conjunto.

**Algoritmo de Hunt (recursivo)**

Se parte de un conjunto de registros de entrenamiento. Si todos pertenecen a la misma clase, se crea un nodo final. Si representan más de una clase, se define una CTA y tantos nodos hijos como resultados posibles haya. Se reparten los registros a cada nodo y se vuelve a aplicar el algoritmo.

Métodos para la CTA

* *Atributo binario:* genera dos posibles resultados.
* *Atributo nominal:* puede tener varios valores sin ordenación. División múltiple o división binaria sucesiva.
* *Atributo ordinal:* atributos que se asocian a una escala ordenada. Puede usarse división múltiple o binaria (mantener la relación de procedencia). Ejemplo: “leve”, “moderado”, “grave”.
* *Atributo continuo:* división binaria, división múltiple.

**Errores de un modelo**

* Error de entrenamiento: registros del conjunto de entrenamiento mal clasificados.
* Error de generalización: registros del conjunto de validación mal clasificados.
* Sobre-entrenamiento: error de entrenamiento muy bajo y de generalización muy alto. Puede ser por errores en los registros, falta de registros representativos en el conjunto de entrenamiento o sobre entrenamiento.

Un buen modelo debe tener bajo error de entrenamiento y de generalización.

**Clasificación en base a reglas**

Clasificación de objetos en base a un conjunto R de reglas “Si… entonces…”.

Calidad de una regla de clasificación

**Cobertura** = Soporte

**Exactitud** = Confidencia

Donde:

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

Propiedad de un conjunto de reglas

Mutuamente exclusivo: como máximo se dispara una regla para cada registro.

Exhaustivo: posee una regla para cada combinación de valores de los atributos todo registro es cubierto por, al menos, una regla.

Excepciones:

* *No exhaustivo:* un registro no tiene cobertura. Se agrega una regla por default para cubrir los casos restantes, clasificándolos en la clase mayoritaria.
* *No mutuamente excluyente:* un registro es cubierto por 2 o más reglas con resultado distinto. Solución: reglas ordenadas por prioridad decreciente.

**Big Data**

Utilizado para describir cantidades masivas de datos cuyo análisis en tecnologías de bases de datos tradicionales tomaría demasiado tiempo y sería muy costoso.

Big Data vs. Data Warehouse

Data Warehouse

* Proceso complejo y costoso.
* Datos estructurados.
* Se conoce el alto valor de negocio de los datos.

Big Data

* Proceso más simple y menos costoso.
* Datos estructurados, no estructurados y semi.
* Incertidumbre sobre calidad y valor de los datos.

Usos de Big Data: análisis de datos en procesos iterativos o exploratorios.

**Plataforma Hadoop**

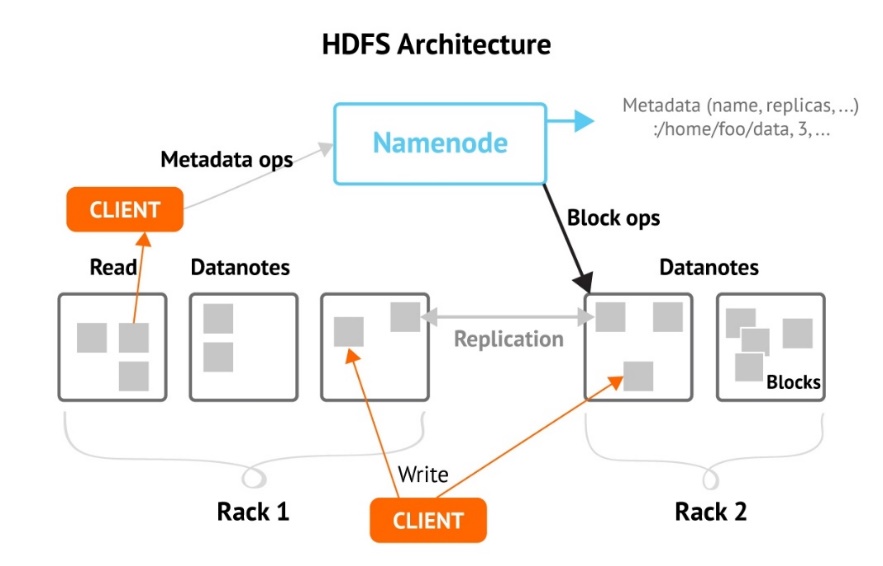
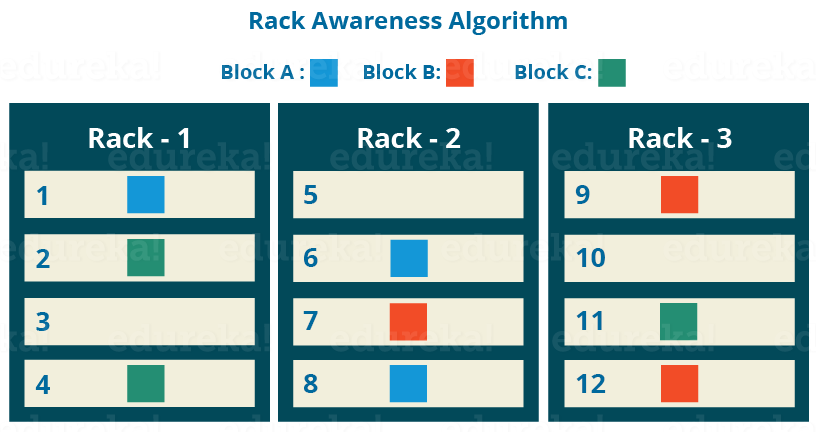
* Diseñado para facilitar la escalabilidad.
* Manejo y recuperación automática de errores.
* Basado en la utilización de hardware de bajo porte.
* Reducción de costos de hardware, software y mantenimiento.

Componentes

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

Hadoop Distributed File System

* Sistema de archivo distribuido, tolerante a fallos y basado en hardware de bajo porte. Los archivos se almacenan en clúster, en forma redundante.
* Brinda rapidez y facilidad de acceso.
* Los datos se dividen en bloques, permitiendo procesamiento paralelo.
* *Componentes:*
  + **Name Node**: administra la metadata y controla 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.
* Simplicidad de programación.
* Gran capacidad de escalabilidad 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 asigna una única clasificación.
  + Shuffle: unifica todos los paquetes mapeados para cada criterio.
  + Reduce: procesa los paquetes unificados para obtener reportes y estadísticas.

