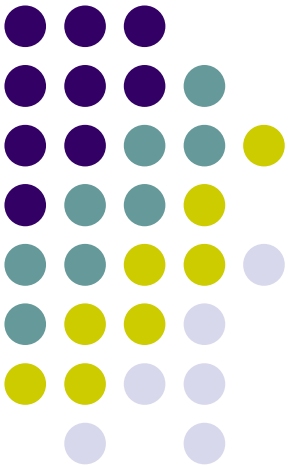
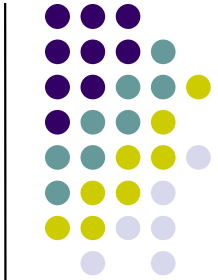
Data Warehouse
Elementos de diseño

Parte 3:

Diseño



Data Warehouse
Diseño



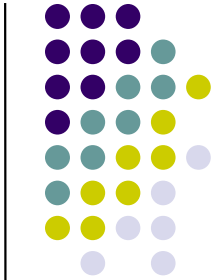
Planificación

- “El 70% del tiempo total dedicado al proyecto se insume en definir el problema y en preparar los datos”.

- “Estime el tiempo necesario, multiplíquelo por dos y agregue una semana de resguardo”.

- Regla 90 – 90: “el primer 90% de la construcción de un sistema absorbe el 90% del tiempo y esfuerzo asignados; el último 10% se lleva el otro 90% del tiempo y esfuerzo asignado”.

Data Warehouse
Diseño

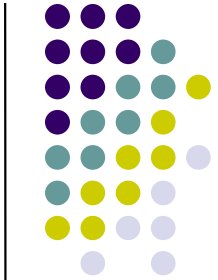


Performance

- Selección y configuración del SGBD que dará soporte al DW
- Elección de estrategias para mantener las estructuras de datos.

La Base de Datos

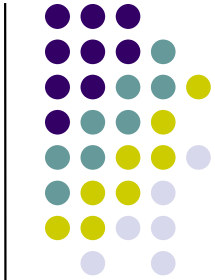
- Seleccionar los tipos de datos, por ejemplo, para valores enteros pequeños conviene utilizar **tinyint** o **smallint** en lugar de **int.** ya que formarán parte de las tablas de Hechos (que son las de mayor volumen) además de que toda clave primaria tiene asociado un índice que la implementa.
- Utilizar Claves Subrogadas. Utilizar técnicas de indexación.
- Utilizar técnicas de particionamiento. Crear diferentes niveles de agregación.
- Utilizar distribución de datos.
- Utilizar técnicas de multiprocesamiento distribuido, con el objetivo de agilizar la obtención de resultados, a través de la realización de procesos en forma concurrente.



Performance

Elección de columnas

- Descartar columnas cuyos valores tengan muy poca variabilidad.
- Descartar columnas que tengan valores diferentes para cada objeto, por ejemplo, el número de documento, cuando se analizan personas.
- En los casos en exista una constelación, intentar embeberla en la dimensión más cercana a los hechos implementando una jerarquía.



Performance

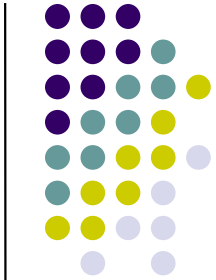
Tratamiento de las columnas

- **Factorizar:** se utiliza para descomponer un valor en dos o más componentes

- **Estandarizar**: se utiliza para ajustar valores a un tipo de formato o norma preestablecida.

- **Codificar**: se utiliza para representar valores a través de las reglas de un código preestablecido. Por ejemplo, en la columna **estado** se pueden codificar sus valores, **0** y **1**, para transformarlos en **Apagado** y **Encendido**

- **Discretizar**: se utiliza para convertir un conjunto continuo de valores en uno discreto. Por ejemplo, en el campo **intensidad** se pueden codificar los valores menores a 100 como **Baja**; los valores mayores a 100 y menores a 500 como **Media**; y los valores mayores a 500 como **Alta**.



Performance

Tratamiento de las relaciones

Evitar mantener en el DW tablas de Dimensiones con relaciones muchos a muchos entre ellas. Las soluciones podrían ser:

- Crear una tabla de relación (que también sería una Dimensión) que tenga las claves de ambas tablas quedando de esta manera 1,n con cada una de ellas para luego vincular la tabla de hechos con esta tabla de relación.
- Agregar las dos claves primarias de las tablas de Dimensiones en la tabla de Hechos.

La única desventaja es en cuanto a los procesos ETL, ya que se aumenta su complejidad y tiempo de proceso.

Data Warehouse
Diseño

Performance

- **Claves naturales o de negocio**
- **Claves subrogadas**
  - Numérico secuencial,
  - NO tienen relación directa con ningún dato ni tienen significado
  - Ocupan menos espacio y son más performantes que las claves naturales, y más aún si éstas son tipo texto.
  - Construcción y mantenimiento de índices más sencillo.
  - El Data Warehouse NO dependerá de la codificación de los Data Sources.
  - Si se modifica el valor de una clave en el Data Source, el DW lo tomará como un nuevo elemento, almacenando diferentes versiones del mismo dato.
  - Permiten la correcta aplicación de técnicas **SCD**.

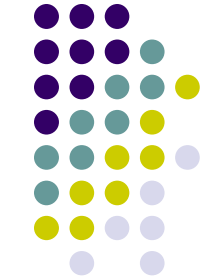
SCD (slowly changing dimensions)

Cuando ocurren cambios en las dimensiones se puede:

- Registrar el **historial** de cambios.
- **Reemplazar** los valores que sean necesarios.

Ralph Kimball planteó tres estrategias a seguir cuando se tratan las SCD: tipo 1, tipo 2 y tipo 3; pero a través de los años se profundizaron las definiciones iniciales e incluyó tipo 4 y tipo 6.

- **SCD Tipo 0**: no tener en cuenta ningún cambio.
- **SCD Tipo 1**: Sobrescribir.
- **SCD Tipo 2**: Añadir fila.
- **SCD Tipo 3**: Añadir columna.
- **SCD Tipo 4**: Historial separado.
- **SCD Tipo 6**: Híbrido.



SCD Tipo 1 - Sobrescribir

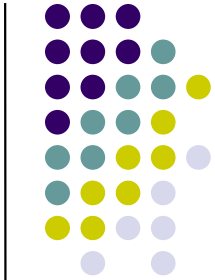
idProducto	Rubro	Tipo	Producto
1	Rubro 1	Tipo 1	Producto 1

Ahora, se supondrá que este producto ha cambiado de **rubro**, y ahora ha pasado a ser *Rubro 2*, entonces se obtendrá lo siguiente:

idProducto	Rubro	Tipo	Producto
1	Rubro 2	Tipo 1	Producto 1

Se usa donde la información histórica no sea importante de mantener (como errores de ortografía).

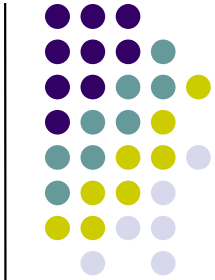
El ejemplo muestra ahora que todos los movimientos realizados de *Producto 1*, que antes pertenecían al *Rubro 1*, ahora pasarán a ser del Rubro 2, lo cual creará inconsistencias en el DW.



SCD Tipo 2 – Añadir fila

Requiere el agregado previo de las columnas:

- fechaInicio: fecha desde que entró en vigencia el registro actual. Por defecto suele utilizarse una fecha muy antigua, ejemplo: 01/01/1000.
- fechaFin: fecha en la cual el registro actual dejó de estar en vigencia. Por defecto suele utilizarse una fecha muy futurista, ejemplo: 01/01/9999.
- version: número secuencial que se incrementa cada nuevo cambio. Por defecto suele comenzar en 1.
- versionActual: especifica si el campo actual es el vigente. Este valor puede ser en caso de ser verdadero: true o 1; y en caso de ser falso: false o 0.



SCD Tipo 2 – Añadir fila

idProducto	CN_Producto	Rubro	Tipo	Producto
1	155	Rubro1	Tipo1	Producto1

Agregado de columnas:

idProducto	CN_Producto	Rubro	Tipo	Producto	FechaInicio	FechaFin	Version	VersionActual
1	155	Rubro1	Tipo1	Producto1	01/01/1000	01/01/9999	1	true

Este producto ha cambiado de Rubro, y ahora ha pasado a ser Rubro2, entonces se obtendrá lo siguiente:

idProducto	CN_Producto	Rubro	Tipo	Producto	FechaInicio	FechaFin	Version	VersionActual
1	155	Rubro1	Tipo1	Producto1	01/01/1000	09/05/2023	1	false
2	155	Rubro2	Tipo1	Producto1	09/05/2023	01/01/9999	2	true

SCD Tipo 2 – Añadir fila

Proceso

idProducto	CN_Producto	Rubro	Tipo	Producto	FechaInicio	FechaFin	Version	VersionActual
1	155	Rubro1	Tipo1	Producto1	01/01/1000	09/05/2023	1	false
2	155	Rubro2	Tipo1	Producto1	09/05/2023	01/01/9999	2	true

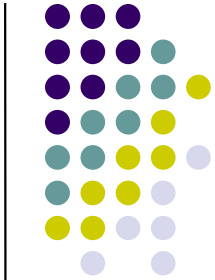
- Se añade una nueva fila con su correspondiente clave subrogada (idProducto = 2) manteniéndose la clave de negocios original (CN_Producto)
- Se registra la modificación (Rubro).
- Se actualizan los valores de FechaInicio y FechaFin, tanto de la fila nueva, como la antigua (la que presentó el cambio).
- Se incrementa en uno el valor del campo Version que posee la fila antigua.
- Se actualizan los valores de VersionActual, tanto de la fila nueva, como la antigua; dejando a la fila nueva como el registro vigente (true).

SCD Tipo 2 – Añadir fila

Proceso

idProducto	CN_Producto	Rubro	Tipo	Producto	FechaInicio	FechaFin	Version	VersionActual
1	155	Rubro1	Tipo1	Producto1	01/01/1000	09/05/2023	1	false
2	155	Rubro2	Tipo1	Producto1	09/05/2023	01/01/9999	2	true

- Permite mantener todos los cambios (ilimitado).
- La referencia de los hechos se hace al idProducto pero que corresponda a la fecha de vigencia y la misma CN_Producto.
- Distintos Hechos pueden corresponder al mismo producto pero en períodos distintos (mismo CN_Producto y distinto idProducto).



SCD Tipo 3 – Añadir columna

Requiere el agregado previo una columna adicional por cada columna cuyos valores se desean mantener en el historial:

idProducto	CN_Producto	Rubro	Tipo	Producto
1	155	Rubro1	Tipo1	Producto1

Para mantener el histórico sobre los datos de la columna Rubro, se agregará la columna RubroAnterior (mismo tipo de datos). Si cambia de Rubro1 a Rubro2 se obtendrá:

idProducto	CN_Producto	Rubro	RubroAnterior	Tipo	Producto
1	155	Rubro2	Rubro1	Tipo1	Producto1

- En la columna RubroAnterior se coloca el valor antiguo *(Rubro1).*
- En la columna Rubro se coloca el valor vigente *(Rubro2)*

Mantiene UN SOLO CAMBIO: *El último.*

SCD Tipo 4 – Historial separado

- Se utiliza en combinación con alguna de las anteriores.
- Se crea una tabla adicional con los detalles de los cambios históricos.

La Dimensión original es:

idProducto	CN_Producto	Rubro	Tipo	Producto
1	155	Rubro1	Tipo1	Producto1

Y suponiendo si se trabaja con SCD Tipo 2 y cambia **Rubro** de Rubro1 a Rubro2 y **Producto** de Producto1 a Producto50 para **CN_producto** = 155, la tabla de dimensión será ahora:

idProducto	CN_Producto	Rubro	Tipo	Producto
2	155	Rubro2	Tipo1	Producto50

Y el historial será:

idProducto	CN_Producto	Rubro	Tipo	Producto	FechaInicio	FechaFin	Version	VersionActual
1	155	Rubro1	Tipo1	Producto1	01/01/1000	09/05/2023	1	false
2	155	Rubro2	Tipo1	Producto50	09/05/2023	01/01/9999	2	true

SCD Tipo 4 – Historial separado

idProducto	CN_Producto	Rubro	Tipo	Producto	FechaInicio	FechaFin	Version	VersionActual
1	155	Rubro1	Tipo1	Producto1	01/01/1000	09/05/2023	1	false
2	155	Rubro2	Tipo1	Producto50	09/05/2023	01/01/9999	2	true

Se asemeja a la manera que se implementan las tablas de auditoría de las bases de datos (registro anterior, registro actual).

Si se implementa SCD Tipo 3 y manteniendo el valor anterior de las columnas **Rubro** y **Producto** cuando se modifican se tendrá mayor control sobre los cambios.

idProducto	CN_Producto	Rubro	Rubro_Anterior	Tipo	Producto	Producto_Anterior	FechaInicio	FechaFin	Version	VersionActual
1	155	Rubro1	Rubro1	Tipo1	Producto1	Producto1	01/01/1000	09/05/2023	1	false
2	155	Rubro2	Rubro1	Tipo1	Producto50	Producto1	09/05/2023	01/01/9999	2	true

SCD Tipo 6 – Híbrido

SCD Tipo **1** + SCD Tipo **2** + SCD Tipo **3** = **SCD Tipo 6**

idProducto	CN_Producto	Rubro	Rubro_Ant	Tipo	Producto	FechaInicio	FechaFin	Version	VersionActual
1	155	Rubro1	Rubro1	Tipo1	Producto1	01/01/1000	01/01/9999	1	true

El **Rubro** y **Rubro_Ant** son iguales. El atributo **VersionActual** indica que este es el registro actual o más reciente para este producto. Cuando el *Producto1* cambia de *Rubro*1 a *Rubro2* se agrega un nuevo registro por SCD Tipo 2 quedando:

idProducto	CN_Producto	Rubro	Rubro_Ant	Tipo	Producto	FechaInicio	FechaFin	Version	VersionActual
1	155	Rubro1	Rubro1	Tipo1	Producto1	01/01/1000	09/05/2023	1	false
2	155	Rubro2	Rubro1	Tipo1	Producto1	09/05/2023	01/01/9999	2	true

Se sobrescribe la información de **VersionActual** como en el SCD Tipo 1. Se crea un nuevo registro para rastrear los cambios como en el SCD Tipo 2, y se almacena el historial en la columna **Rubro_Ant** que corresponde al SCD Tipo3.

Dimensiones y Hechos especiales

1. **Dimensiones realizadoras de roles**

2. **Dimensiones no deseadas**

Dimensión tipoDocumento		Dimensión sexo	
idTipodoc	nombreTipoDoc	idSexo	nombreSexo
1	DNI	1	MASCULINO
2	CEDULA	2	FEMENINO
3	PASAPORTE		

Dimensión tipoVarios		
idtipoVarios	tipoDocu	sexo
1	DNI	MASCULINO
2	DNI	FEMENINO
3	CEDULA	MASCULINO
4	CEDULA	FEMENINO
5	PASAPORTE	MASCULINO
6	PASAPORTE	FEMENINO

3. **Dimensiones degeneradas**

4. **Tablas de hechos sin hechos**