# ¿Como planificarías e implementarías ETL?

Realizar el diseño lógico, comprender la arquitectura DW/BI y mapear la fuente.

El primer paso es visualizar e identificar las fuentes y destinos, y en base a esto ver las transformaciones necesarias y seleccionar las herramientas ETL. Se mapea el flujo de datos, cargando las tablas dimensiones y las tablas hechos, se procesan ambas tablas y se agrega la carga de tabas. Finalmente se realizan las operaciones y se automatiza.

# Las V’s de Big Data

* Veracidad: los datos deberán aportar hechos y ser precisos
* Volumen: habrá una gran cantidad de datos
* Valor: aportarán valor a la empresa
* Variedad: hay datos de varias clases
* Velocidad: deberán ser relativamente rápidos de procesar

# Operaciones OLAP

* Aumentar dimensiones (“alejar” el cubo): ROLL UP
* Disminuir dimensiones (“acercar” el cubo): DRILL DOWN
* Pasar de 3D a 2D (generar un plano): SLICE
* Pasar a un subcubo: DICE
* Representar un plano y rotarlo: PIVOT

# SOM

Self-Organizing Map. Está basado en redes neuronales, y compiten entre ellos para representar mejor los datos. Se define el número de nodos y el nivel de detalle, con pesos inicializados aleatoriamente. Se comparan las entradas con los nodos, y se ajustan los pesos.

Tiene un factor de aprendizaje que determina lo rápido que converge la red.

# GMM

Ajustar los datos a un modelo matemático de superposiciones de varias distribuciones. El número de clústeres se obtiene usando una aproximación bayesiana. Se selecciona el número de clústeres y se inicializan los parámetros de la distribución gaussiana.

# Esquema Estrella

Se usan por ejemplo en un esquema de ventas de una compañía. Consultas simples, reportes fáciles, cubos OLAP.

Puede haber facturas en las tablas hechos, datos de clientes, productos, etc.

# Modelos Multidimensionales

* Modelo Estrella: descrito arriba.
* Modelo Copo de Nieve: normalizar tablas dimensiones con respecto al modelo de estrella. Menos espacio ocupado, pero menor rendimiento.
* Modelo Constelación: múltiples tablas de hechos compartiendo tablas de dimensiones.

# Agregación y Clasificación. Regresión.

La agregación es un método no supervisado, mientras que la clasificación es supervisado. Es decir, para la agregación los resultados se procesan sin tener que entrenarlos antes, y no es necesario generar un modelo de antemano, mientras que la clasificación tendrá que tener unos datos previos de entrenamiento (supervisado).

La regresión es un tipo de clasificación cuando se predice una variable continua.

# Random Forest

Combinación de varios modelos predictivos de diferentes algoritmos de clasificación. Generalmente es un ensamblado de múltiples árboles de decisión.

# K-NN y K-Means

K-NN es un método de clasificación y K-Means es un método de agregación. K-NN es bastante similar a K-Means.

# ETL vs ELT

ETL tiene un gran número de herramientas, procesos bien definidos y gran tradición, pero es poco flexible y con cambios costosos en tiempo y dinero. No se soportan los lagos de datos.

En ELT los datos se cargan directamente en el DW y están disponibles rápido, por lo que es compatible con data lakes. Es sencillo de implementar y escalable, compatible con la nube. Apropiado para transformaciones que requieren cambios frecuentes.

# Rendimiento de una Base de Datos

* Hardware (más discos, más memoria, procesador, etc.)
* Ajustar el Sistema Operativo (filesystem y kernel)
* Parámetros y diseño d ela BD: buffer apropiado, ajuste automático, localización de logs en otro disco
* Middleware: drivers, conexiones, caching, etc.
* Aplicaciones que modifiquen el rendimiento

Ajustes de software, como mal diseño, SQL malo, o malas prácticas con muchas transacciones, no uso de vistas materializadas, etc.

# DBSCAN

Método de agrupamiento que trabaja con el concepto de un punto que debe ser alcanzable por densidad. Tres tipos de puntos: core, borde, ruido.

# CRSIP-DM

Estándar de la industria para la minería de datos. Se basa en un proceso cíclico en que se parte de un análisis y entendimiento del negocio, entiendo los datos se pueden preparar y modelar, para finalmente evaluar y poner en práctica.

# Bootstrapping y Bagging

Bootstrapping: remuestrear los datos de manera que puedan repetirse en la muestra.

Bagging: usar muchas veces un conjunto de entrenamiento usando bootstrapping, combinando las salidas por promedio (regresión) o mayoría (clasificación).

# AHP

Proceso Analítico Hierárquico, comparaciones a pares entre atributos para poder evaluar cuantitativamente una serie de alternativas. Una escala de importancia con puntuaciones entre varios atributos a pares, de tal forma que se puntué en base a esas ponderaciones futuras características obteniendo una puntuación global.

# Verdadero Falso

* Normalizar reduce el número de claves foráneas: F
* K-Means funciona mejor en grupos más separados: V
* Maldición de la dimensionalidad: ¿?
* Puede ser que una tabla de hechos no contenga ninguna medida: F
* Más contención significa aguanta más caudal de datos: F
* Indexar puede producir un bloqueo: V