17/11/2023

ETL

Son la siglas de extract, transform y load.

Ventajas

* Recopilar, transformar y consolidar datos de forma automática
* Facilita el trabajo con datos complejos: estructura de datos compleajs a simples
* Reduce errores asociados al factor humano
* Aumenta el ROI (Return of investment)

Conceptos

* Extracción
* Transformación
* Carga (Load): Volcar datos en una estructura que nos permita trabajar de forma comoda en un BI, ML, Data análisis. Usualmente termina en un data warehouse

Componentes

Transformaciones

* Reformateo de datos
* Convesion de unidades
* Agregación de columnas
* Traduccion

Data Warehouse

* Cloud
  + Amazon Redshift
  + Azure Synapse Analytics
  + BigQuery
* On premises:
  + RDBMS

Diseño

* Funcional: requisito de la ETL
* No funcional: requisitios de arquitectura y herramientas a utilizar

Decisiones a tomar

* SW
* Tecnologías de las fuentes
* Conociemitno de los datos
* Donde se gurada la info
* Como se representan los datos. Visualización (a través de BI…)

Arquitectura

ETL son tres elementos principales (extracción, transformación y carga). Muy recomendable aumentar a estos 6:

* Profiling
* Extraction
* Cleasnsing
* Transformation
* Loading
* Monitoring

Casos de uso

* Solamente ingesta: datos limpios, listos para explotar
* Ingesta + Carga:
* Ingesta + Tranformacion + Carga

Otras acciones

* Migrar datos de una aplicación a otra
* Replicar datos (alternativa a copia de seguridad)
* Sincronizar datos con BI

Resumiendo

* Mover datos desde múltiples fuentes
* Reformatearlos, limpiarlos
* Cargarlos en
  + Otra base de datos
  + Datamart
  + Datawarehouse
  + Datalake

Profiling

Examinar los datos sin procesar.

* Descubirir estructuras: rango de valores aceptado (utilizando la varianza o variabilidad)
* Descubrimientos de contenido: Que información contiene cada registro y que se hace con los registros que faltan
* Descubrimiento de relaciones

Extracción

Pentaho

* TRanformacion
* Job: monitorización de tranformaciones

Comandos en cmd

* Pan: ejectuar transformaciones
* Kitchen: ejectuar un job

23/11/2023

Transformación

Guardar siempre una copia de los datos en raw.

Tener en cuenta los siguientes pasos

* Normalización: qué datos son importantes y cómo se deben guardar
* Eliminar duplicados
* Verificación: comprobaciones automatizadas de que la información existe. Generar alarmas si alguno de los pasos falla
* Clasificaición: agrupar y clasificar los datos en bruto

Un proceso de transformación

* Debe: transformar los datos para mejorarlos
* No debe: crear información o duplicar, eliminar información relevante

Mapeo de datos

Útil cuando se tiene que migrar datos o integrar datos. Pasar de provincia a código, etc..

Limpieza de datos

Eliminar los datos nulos o no validos

Otras tranformaciones:

Cambios de codificación (unificar criterios)