INTEGRACION DE DATOS–CURSO 23/24
PENTAHO PDI (KETTLE)

TAREFA B

Considera as dimensións incluídas no voso Datamart. Queremos desenvolver dúas transformacións en Pentaho PDI para actualizar a información das dimensións **non temporais** deseñadas na **tarefa A**, cunha estratexia *insert else update (IeU)* e con actualización SCD 1 de Kimball.

Pasos:

1. Realiza a **caracterización** (***profiling***) das táboas operacionais. ¿Existen posibles problemas (valores nulos, integridade referencial…) que poidan afectar á carga do Datamart?. Lembra eses problemas (se os hai) cando realices a tarefa C.
2. Selecciona o método de **control de cambios** que despois aplicarás no teu proceso ETL. Se fose o caso (por exemplo, se escolles a opción de columnas de auditoría), realiza todos os cambios que necesites na base de datos operacional.
   AVISO: Debes seleccionar un ÚNICO método/variante CDC. Non está permitido combinar no traballo varios métodos/variantes diferentes.
3. Implementa unha **transformación** en Pentaho PDI para cada unha das túas dimensión.
   NOTA: Lembra que **NON DEBES** implementar unha transformación para cargar a dimensión temporal. O contido da táboa dimensional temporal será precargado no script de creaciòn da BD do Datamart producido durante a execución da tarefa A.

Condicións básicas:

- As transformacións deben estar deseñadas para ser executadas coa **periodicidade** (diaria, semanal, mensual, trimestral...) que corresponda ao teu Datamart.[1]
- Debes implementar o método de **control de cambios** elixido por ti para cargar, nas táboas dimensionais, **unicamente filas novas e/ou modificadas**.[2]
- MOI IMPORTANTE: o teu método CDC debe permitir SEGREGAR as filas novas e as filas modificadas, que serán procesadas POR SEPARADO na túa transformación. **Se non o fas así, esta tarefa non será avaliada.**
- Define nas transformacións todos os parámetros de entrada que necesites.[3]
- As táboas dimensionais deben ser actualizadas de acordo coa definición de **SCD tipo I** de Kimball

---

1 No exemplo da tarefa A, a periodicidade sería mensual. Co fin de cada mes recuperaríamos os cambios producidos nos datos de localidades e axentes de ventas producidos no dito mes.

2 No exemplo da tarefa A, unha transformación executada para actualizar xaneiro de 2019 debería recuperar unicamente filas con data de alta ou modificación comprendida entre o 01/01/2019 e o 31/01/2019

3 Por exemplo, un parámetro MES_CARGA (se estamos actualizando xaneiro de 2020, fixamos MES_CARGA a '01/2020'); ou un par de parámetros DATAINI e DATAFIN (se estamos actualizando xaneiro de 2020, fixamos DATAINI a '01/01/2020' e DATAFIN a 31/01/2020)

    - As filas novas serán inseridas na táboa dimensional, cun novo valor para a clave subrogada.[4]
    - As filas modificadas serán actualizadas sobrescribindo a información existente previamente na táboa dimensional.[5]
- Utiliza o ficheiro csv deseñado na tarefa A naquelas transformacións nas que sexa necesario.[6]
- Incorpora na transformación calquera control que consideres oportuno por algún problema detectado no perfilado de datos.

Información adicional

Dispoñedes de información e exemplos sobre o uso das tarefas dispoñibles en Pentaho PDI no seguinte enderezo:

Pentaho Data Integration Steps - Pentaho Data Integration - Pentaho Community Wiki (atlassian.net)

En particular, podedes atopar interesantes:

- *Table input (info)*
- *Tex file input (info)*
- *Stream lookup (info)*
- *Combination lookup/update (info)*
- *Update (info)*

Revisa tamén as diapositivas “Pentaho lookups” dispoñibles no Moodle.

4 No exemplo da tarefa A, novos axentes de ventas serían dados de alta na táboa dimensional asignándolles un novo subrogado.

5 Os datos dun axente de ventas dado previamente de alta na táboa dimensional serán sobrescritos, respectando o subrogado que lle foi asignado anteriormente.

6 Por exemplo, necesitariamos utilizar necesariamente o ficheiro csv na transformación de carga da dimensión *Localidade*