Práctica 2: Procesos ETL en un Data Mart

Almacenes y Minería de Datos
Curso 2020/2021

Grado en Ingeniería en Informática
Dpto. de Informática e Ingeniería de Sistemas
Universidad de Zaragoza
Escuela de Ingeniería y Arquitectura

Sergio Ilarri y Raquel Trillo

30 de octubre de 2020

1. Objetivos / tareas a realizar

En esta práctica se diseñarán e implementarán los mecanismos necesarios para la carga de datos del data mart implementado en Oracle en la práctica anterior. En mayor detalle, los objetivos de esta práctica son:

- Modificar el diseño de la base de datos analítica (OLAP) realizada en la práctica anterior para incluir información acerca de las ciudades donde se encuentran los aeropuertos (si no se había incluido ya en el diseño original).
- Ajustar, si es necesario, el diseño previo en función de los datos realmente disponibles y justificando los posibles cambios.
- Realizar la carga de datos de la base de datos analítica (OLAP) implementada en la práctica anterior a partir de información pública disponible en la Web.

2. Contexto del problema

Se pretende realizar la carga de datos del data mart que manipula información de vuelos comerciales en Estados Unidos implementado en la práctica anterior. En primer lugar se requiere que se almacene mayor cantidad de información acerca de las ciudades donde se encuentran los aeropuertos. Al menos se deben considerar los siguientes datos de cada ciudad: nombre, número de habitantes de dicha ciudad y la zona horaria ("time zone").

Se considerarán como fuentes de información para poblar el data mart al menos los sitios web *GeoNames* (`http://www.geonames.org/`, `http://download.geonames.org/export/dump/`) y *RITA* (`http://www.transtats.bts.gov`). GeoNames es una base de datos geográfica y RITA (*Research and Innovative*

Technology Administration) proporciona información de vuelos comerciales en Estados Unidos. Salvo que se encuentren problemas técnicos insalvables, se deberán cargar al menos los datos de vuelos realizados durante un mes del año (ya transcurrido) y (en la medida de lo posible) es muy recomendable realizar pruebas con volúmenes de datos mayores y analizar el rendimiento (tiempo de carga, etc.) y los posibles problemas que aparezcan. Hay que tener en cuenta que algunos datos que pueden ser necesarios para poblar la base de datos se pueden obtener de forma bastante directa, pero otros (por ejemplo, los referentes a los tipos de modelos de aviones) pueden requerir más ingenio o incluso plantear dificultades que lleven a tener que realizar suposiciones o simplificaciones. Además, con respecto a las ciudades en Estados Unidos hay que tener en cuenta que existen diversas ciudades que tienen el mismo nombre.

Para el diseño de los procesos ETL, se pueden considerar diversas opciones. Por ejemplo, una opción podría suponer la utilización y procesamiento de forma más o menos directa de ficheros Excel o ficheros de texto diseñando un programa adecuado. Otra opción puede ser la utilización de herramientas más especializadas, como *Pentaho Data Integration*[1], *Talend*[2], KNIME Software (`https://www.knime.com/`) u otras. Preferentemente se utilizará alguna herramienta específica (como Pentaho Data Integration) para la realización de la práctica. Se valorará positivamente la utilización de al menos dos herramientas o mecanismos diferentes para desarrollar los procesos ETL. Aunque se pueden utilizar los recursos disponibles en la Universidad de Zaragoza, se recomienda principalmente utilizar una máquina virtual, donde se dispondrá de una mayor libertad para probar las herramientas deseadas (pudiendo por ejemplo instalar las últimas versiones o realizar posibles ajustes que sean necesarios), así como una mayor flexibilidad para resolver posibles problemas.

Nótese que **un aspecto fundamental de la práctica es investigar los conjuntos de datos, consultar documentación, y experimentar y probar distintas estrategias para implementar los procesos ETL, documentando adecuadamente el trabajo realizado**. En la memoria de la práctica debe describirse con claridad y en detalle las estrategias probadas y su diseño. Se valorará especialmente la capacidad mostrada para resolver problemas de forma autónoma, la iniciativa para probar y explorar soluciones y herramientas, y la descripción detallada de los procesos seguidos en la resolución de la práctica (**se espera una descripción lo suficientemente detallada como para poder reproducir fácilmente el proceso seguido**).

Se considera que se va a realizar una única carga de datos, por lo que no es preciso desarrollar técnicas de carga incremental. No obstante, es conveniente analizar en la memoria de la práctica cómo se gestionaría una situación diferente y, si se considera posible, realizar alguna prueba demostrativa a pequeña escala.

3. Máquinas virtuales

En las notas de apoyo para la realización de las prácticas disponibles en Moodle se proporciona información acerca del tipo de máquinas virtuales que se pueden utilizar en la práctica, así como otra información de interés y algunos consejos.

[1] `https://sourceforge.net/projects/pentaho/files/Data Integration/`
[2] `https://www.talend.com/download/`

4. Entrega de la práctica

La práctica se realizará en equipos de tres personas (salvo que existan problemas logísticos, que deberán comentarse y resolverse previamente con el profesor). Al finalizar la práctica se debe entregar un fichero .zip denominado *pN_NIP1_NIP2_NIP3.zip* (donde NIP1, NIP2 y NIP3 son los NIP de los autores de la práctica, con NIP1 $<$ NIP2 $<$ NIP3, y en pN la N representa el número de práctica) con el siguiente contenido:

1. Un *fichero PDF denominado* **informe.pdf**, que contendrá la memoria de prácticas (respuestas a las cuestiones planteadas, esquemas y código desarrollados, etc., según el caso). En la portada de dicho documento se indicarán los componentes del grupo de prácticas (nombre y apellidos y NIP de cada uno) y el número de práctica. El documento contendrá todo lo necesario para la evaluación de las prácticas. El informe de prácticas debe contener al final un apartado de conclusiones personales que incluirá, entre otras cosas, información sobre el tiempo invertido por cada miembro del grupo de prácticas en la realización de la misma.

2. Todos los fuentes, proyectos, ficheros de configuración y programas de prueba desarrollados, si es el caso, en una carpeta ***fuentes***.

Al descomprimir el fichero .zip se deben extraer los ficheros y directorios necesarios en el directorio *pN_NIP1_NIP2_NIP3*. Es importante seguir las convenciones de nombrado y la estructura de ficheros y directorios descrita.

Para la entrega del fichero .zip, se utilizará Moodle del Anillo Digital Docente de la Universidad de Zaragoza. La fecha límite de entrega es el día anterior al de la siguiente sesión de prácticas en el laboratorio a las 23:59.

La práctica entregada debe contener, además de la memoria, todos los ficheros fuentes y proyectos realizados para su resolución.

Se recomienda incluir en la memoria, entre otros, apartados independientes para describir las fuentes de datos utilizadas, comentar las herramientas probadas y seleccionadas, describir el proceso ETL, reflexionar cómo podría realizarse el proceso de carga incremental, e indicar los problemas encontrados y las soluciones adoptadas.

5. Recomendaciones

Una vez realizadas las prácticas y entregadas, cada grupo debe presentárselas al profesorado de prácticas en la siguiente sesión de prácticas. Al realizar la presentación el profesorado le formulará cuestiones sobre las decisiones que ha tomado. La práctica debe entregarse en los términos indicados anteriormente y debe ser original (no debe haber sido copiada). Se considerarán los siguientes aspectos:

- La consecución de los objetivos planteados y la adecuada realización de las tareas correspondientes.
- La justificación de las afirmaciones incluidas en el informe.
- La adecuada inclusión de referencias a fuentes de información, si procede.

- La metodología para llevar a cabo la búsqueda de información necesaria para la realización de la práctica.
- La creatividad para la búsqueda de soluciones apropiadas y el correcto análisis de alternativas.
- La estructura y presentación del informe elaborado (completitud del trabajo, precisión, explicaciones adecuadas y completas, justificación de las decisiones tomadas, referencias adecuadas, coherencia de discurso, ausencia de errores tipográficos y ortográficos, etc.).

Hay que asegurarse de que, en caso de que la práctica incluya algún tipo de desarrollo, funciona correctamente en los ordenadores del laboratorio o puede probarse fácilmente sobre una máquina virtual base proporcionada por el profesorado. También es importante someter código limpio (donde se ha evitado introducir mensajes de depuración y comentarios que no proporcionan información relevante).

Es importante extraer conclusiones: resultados obtenidos, indagaciones realizadas, diferencias observadas entre las distintas alternativas y/o herramientas, dificultades encontradas, valoración de los aprendizajes o mejoras de habilidades conseguidas, opinión personal, etc. Se recomienda destacar especialmente aquellos aspectos que muestren la capacidad de indagación, búsqueda de información, autonomía, y curiosidad personal que se hayan desarrollado.