Almacenes y Minería de Datos.
Práctica 1.

Profesora: Dra. Amparo López Gaona
alg@ciencias.unam.mx
Profesor: M. en I. Gerardo Avilés Rosas
gar@ciencias.unam.mx
Laboratorio: Lic. Carlos Augusto Escalona Navarro
caen@ciencias.unam.mx

7 de marzo de 2021

Se dan a conocer especificaciones de entrega para la práctica 1.

1. ¿Qué es un proceso ETL?

Los procesos ETL son una parte de la integración de datos que constan de tres fases: **extracción**, **transformación** y **carga**.

1.1. Extracción

Para llevar a cabo el proceso de extracción, se sugiere realizar los siguientes pasos:

- Extraer los datos desde los sistemas de origen.
- Analizar los datos extraídos.
- Interpretar los datos para verificar que estos cumplen la pauta o estructura que se esperaba. Si no fuese así, los datos deberían ser rechazados.
- Convertir los datos a un formato preparado para iniciar el proceso de transformación

Es importante tener en cuenta que durante el proceso de extracción sería el exigir que esta tarea cause un impacto mínimo en el sistema de origen. Este requisito se basa en la práctica ya que, si los datos a extraer son muchos, el sistema de origen se podría ralentizar e incluso colapsar, provocando que no pudiera volver a ser utilizado con normalidad para su uso cotidiano.

1.2. Transformación

En esta fase se aplican una series de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Estas directrices pueden ser declarativas, pueden basarse en excepciones o restricciones pero, para potenciar su pragmatismo y eficacia, hay que asegurarse de que sean:

- Declarativas.
- Independientes.
- Claras.
- Con una finalidad útil para el negocio.

1.3. Carga

Aquí los datos procedentes de la fase anterior son cargados en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes.

Existen dos formas básicas de desarrollar el proceso de carga:

- **Acumulación simple**: esta manera de cargar los datos consiste en realizar un resumen de todas las transacciones comprendidas en el período de tiempo seleccionado y transportar el resultado como una única transacción hacia el **data warehouse**, almacenando un valor calculado que consistirá típicamente en una suma o un promedio de la magnitud considerada. **Es la forma más sencilla y común de llevar a cabo el proceso de carga.**
- **Rolling**: este proceso sería el más recomendable en los casos en que se busque mantener varios niveles de **granularidad**. Para ello se almacena información resumida a distintos niveles, correspondientes a distintas agrupaciones de la unidad de tiempo o diferentes niveles jerárquicos en alguna o varias de las **dimensiones** de la magnitud almacenada (por ejemplo, totales diarios, totales semanales, totales mensuales, etc.).

Hay que tener en cuenta que esta fase interactúa directamente con la base de datos de destino y, por eso, al realizar esta operación se aplicarán todas las restricciones que se hayan definido en ésta. Si **están bien definidas**, **la calidad de los datos** en el proceso ETL estará garantizada.

2. PostgreSQL

Un SMBD es un Sistema Manejador de Bases de Datos en el cual existe una conexión o relación entre datos. Además de ser un sistema donde se almacena información en forma de registros, se puede realizar modificaciones y borrados sobre los mismos datos. Estos sistemas también proporcionan métodos para mantener la **integridad de**

los datos, para administrar el acceso de usuarios a los datos y para recuperar la información si el sistema se corrompe. Los SMBD están diseñados para gestionar grandes cantidades de información.

PostgreSQL es un SMBD relacional desarrollado por PostgreSQL Global Development Group, este se enfoca para un uso gratuito y libre, por lo cual la comunidad de este SMBD es muy grande, sus principales características son el soporte de transacciones, estabilidad, escalabilidad, entre otras. Para esta **práctica del curso** se utilizará **PostgreSQL en su versión 10.14**.

2.1. Extracción de datos en PostgreSQL

Para esta parte utilizaremos el comando **COPY** de PostgreSQL, este comando nos permitirá copiar tablas o consultas a un archivo **CSV**.

El delimitador **';'** es el carácter ASCII único que separa las columnas de cada fila (línea) del archivo. El valor predeterminado en modo texto es un carácter de tabulación o una coma en el modo CSV.

Incluimos la palabra **HEADER** para especificar que el archivo creado contenga una línea de cabecera con los nombres de las columnas de las tablas en el archivo.

Para mayor información sobre el uso del comando se puede consultar la siguiente liga `https://www.postgresql.org/docs/10/sql-copy.html`.

2.2. Transformación de datos en PostgreSQL

Para la parte de transformación de datos en **SQL**, para los campos de tipo se pueden utilizar todas las funciones que vienen integradas en el SMBD para dar o ajustar el formato de los datos de nuestras fuentes de orígenes, para tener una mejor idea de todas estas funciones que se incluyen en el SMBD podemos consultar la siguiente liga `https://www.postgresql.org/docs/10/functions.html`.

2.3. Carga de datos en PostgreSQL

Para la carga de datos volveremos a utilizar el comando **COPY** con una pequeña variación, debido a que ahora de los CSV generados en el proceso de extracción y transformación cargaremos estos datos en nuestro fuente destino.

3. Actividad

Para esta práctica utilizaremos nuestro SMBD PostgreSQL, para crear dos bases de datos las cuales se llamaran Farmacia y Pharmacy, después de haber creado las bases

de datos se deberá restaurar los archivos backup Farmacia.bacckup y Pharmacy.backup respectivamente.

Posterior a esto se deberá realizar un análisis de cada una de las tablas que se encuentran en cada uno de los esquemas para ir encontrando las similitudes entre sí y posterior a este análisis podamos empezar a definir el diseño de nuestro esquema destino, el cual contendrá **la información unificada de ambos esquemas**.

Para esta primera práctica solo nos vamos a enfocar en la información de los **clientes, empleados y sucursales** de ambos esquemas. Es importante que tomes en cuenta que esta información se puede encontrar en diferentes tablas según el diseño de cada esquema.

Una vez que tengamos localizada la información en común de ambos esquemas y sepamos cuales son las tablas involucradas, procederemos a empezar nuestro proceso de ETL haciendo uso de nuestros conocimientos en **SQL**, por lo cual todas las transformaciones necesarias para ajustar los campos de tus tablas se deberán realizar por medio de funciones de SQL, posterior a esto deberás generar tus archivos CSV los cuales tendrás que cargar en tu esquema destino.

4. Entregables

Para esta práctica se entregar lo siguiente:

- Un archivo llamado **diseño.pdf** con el análisis de las tablas involucradas para las entidades de **clientes, empleados y sucursales**, este archivo debe de contener el diagrama relacional de tu esquema destino de estas 3 entidades.
- Un archivo llamado **ddl_destino.sql**, el cual contendrá la definición en **SQL** de tus tablas del esquema destino. No olvides que es importante que estas tablas tengan definidas al menos la integridad de entidad y de dominio y tampoco olvides agregar los comentarios en cada una de las columnas de tus tablas.
- Se deberán generar al menos 3 archivos CSV, los cuales contendrán la información de tus datos resultado de los paso de la extracción y transformación. El nombre de estos archivos deberán ser **clientes.csv, empleados.csv y sucursales.csv**, no olvides incluir **tus archivos sql** utilizados para la generación de estos datos.
- Un archivo llamado **carga.sql**, el cual contendrá las instrucciones en SQL, para poder cargar los archivos generados en el punto anterior.

Fecha de entrega: jueves 18 de marzo de 2021 antes de las 09:59 hrs.