Informe Descripción Proceso de Carga y Proceso de Refresco

Universidad del Valle - KDD

COLMOVIL

Versión 1.0
15/10/14

Realizado por:

Luis Felipe Murillo.
John Andrés Medina.
Juan Manuel Olaya.
Esteban Antonio Llanos.

Tabla de Contenido

DESCRIPCIÓN PROCESO DE CARGA

Para el proceso de carga fue necesario realizar primero el montaje de la base de datos relacional, pero para ello debimos realizar una limpieza y adaptación previa de los datos - de la base de datos relacional - con el fin de llevar a cabo esta primera etapa, puesto que si existían problemas de integridad referencial en la información, no sería posible montar la base de datos para así aplicarle el proceso de ETL.

Inicialmente creamos el esquema y las tablas haciendo uso del archivo SQL dado que contenía la información de la Base de Datos Relacional. Para esta etapa no se tomaron demasiadas consideraciones, solo se cargo la base de datos relacional, usando directamente el archivo **"colmovil.sql"**. Una vez hecho esto se procedió a realizar la etapa de inserción de los datos, pero en esta parte se presentaron inconvenientes puesto que en algunas tablas se tenían llaves foráneas de registros inexistentes.

Adicionalmente se presentó la situación de tener dos veces los mismos datos de contratos en el archivo **"contratos_sim_card.csv"** y **"contratos_sim_carddef.csv"** con la única diferencia de que el id del contrato en el archivo **"contratos_sim_carddef.csv"** estaba dado por una secuencia que aumentaba de tres en tres mientras que en el archivo **"contratos_sim_card.csv"** se realizaba el incremento de uno en uno.
Para esto debimos tomar el archivo **"contrato_sim_carddef.csv"** e ignorar el otro ya que las referencias desde otras tablas se hacían a usuarios con el id incrementando de 3 en 3.
Para solucionar el problema de las llaves foráneas sin referencia en los datos de contrato, se modificó el valor del id de los clientes, en los registros donde fallaba la referencia, con base en los clientes existentes (10000 registros de clientes, con id's entre 1 y 10000), creando un id de cliente aleatorio entre 1 y 10000**.**

Una vez resueltos los problemas de integridad referencial que no permitían el montaje de la base de datos relacional, conformamos el script de creación de la bodega en el archivo **"dwh_colmovil.sql"** adjunto a este informe. En él se consideraron los aspectos a tener en cuenta para dar respuesta a los interrogantes de cada proceso de negocios. Por lo tanto solo se seleccionó la información relevante.
Durante la producción de la bodega se tomaron las debidas precauciones en cuanto a la no nulidad de los atributos, los valores por defecto que tomarían los campos vacíos y los demás aspectos para garantizar la seguridad en la información contenida dentro de la bodega de datos.

Los Scripts para la carga de las dimensiones: Oficina, Equipo, Plan_Voz, Plan_Datos, Cliente y Sim_Card se realizaron en java puesto que la cantidad de registros contenidos en estas dimensiones permitían realizar el proceso de carga para estas dimensiones en un periodo corto de tiempo (con un máximo de tiempo de unos 8 minutos aproximadamente). En estos scripts se llevó a cabo el

proceso de extracción, transformación y carga de cada una de las dimensiones mencionadas.

Para la carga de las dimensiones: Tiempo y Fecha se usó el mismo Script de creación de la bodega. Vale aclarar que las dimensiones Fecha y Tiempo se cargaron a partir de la información interna de Postgres, por lo cual no requirieron ningún proceso de limpieza.

Inicialmente se consideró la opción de cargar las dimensiones localización y demografía a partir del Script de la bodega de datos pero finalmente se decidió que era más conveniente manejar estas dimensiones desde Scripts independientes puesto que si se mantienen en el Script de creación de la bodega, se perdería información histórica al momento de llevar a cabo un eventual proceso de refresco.

Para la carga de las tablas de hechos se presentaron algunos inconvenientes en ciertos procesos de negocios. En llamadas en particular, se tenían más de medio millón de registros, por lo cual hacer uso de un Script realizado en java no era muy eficiente, ya que todo el proceso de ejecución del Script tomaría horas en llevarse a cabo. Vista esta situación, fue necesario realizar un nuevo Script en sql ya que los tiempos de ejecución se reducirían en comparación con la implementación hecha en java (la cual tardaba alrededor de un minuto para cargar tan sólo 100 registros).

Al igual que con las dimensiones, algunas de las tablas de hechos se cargaron mediante la implementación en java, esto se hizo con tablas de hechos que no contaran con muchos registros y con lo cual no se presentara el problema de tiempos de ejecución.

Los Scripts de ejecución de las dimensiones y las tablas de hechos están adjuntos al presente informe.

DESCRIPCIÓN PROCESO DE REFRESCO

Dado que el proceso de refresco consiste en la carga de nueva información al interior de nuestra bodega de datos y en vista de que nosotros solo ejecutaremos la carga inicial de la misma durante el desarrollo de este proyecto, podemos concluir que no llevaremos a cabo este proceso, por lo cual no nos es posible dar una descripción de este.

En caso de que se nos pidiese llevar a cabo un proceso de refresco sería necesario ejecutar los Scripts de carga nuevamente sobre la base de datos relacional con la nueva información y de esta manera cargar los datos correspondientes en la bodega de datos.

Adicionalmente es necesario corroborar si en la base de datos relacional se llevaron a cabo cambios estructurales en el manejo de la información, es decir, si fueron alteradas las tablas de la base de datos, ya sea por la creación de nuevas columnas, así como una posible creación de nuevas tablas que afecten la interpretación de nuestro proceso de negocios. Si este fuese el caso sería necesario modificar tanto la bodega de datos como los Scripts de carga que se ejecuten sobre la bodega.

BIBLIOGRAFÍA

- **http://www.cs.uoi.gr/~pvassil/downloads/ETL/SHORT_DESCR/08SpringerEncyclopedia_draft.pdf**

[Información acerca el Proceso de Refresco]

- **http://www.dataself.com/wiki/DataSelf_ETL_Glossary#Refreshing_Data**

[Información Proceso de Refresco]