UNIVERSIDAD DE ZARAGOZA

Doc: Sergio Ilarri

Asignatura: Manipulación y análisis de grandes volúmenes de datos.

Memoria de de prácticas I - II

Elaborada por:

1. Rodolfo Palacio Abrego - 792168
2. Gian Marco Berni - 867084
3. Norman Bellorin - 794866.

Fecha de entrega: 27/04/2022

Introducción

El presente documento explica detalladamente el procedimiento seguido para la elaboración de un DataMart, el cual se desarrolla sobre el ámbito de películas de cine. En el documento estudiaremos cómo se llevó a cabo la creación de la base de datos, la selección de la tabla de hechos y sus dimensiones, así como el procedimiento seguido que nos llevó a dichas decisiones.

Se explicarán también detalles técnicos de tecnologías utilizadas como gestores de base de datos, lenguajes de programación utilizados para la trata de procesamiento de datos, etc.

También se hablará del procedimiento ejecutado para la población del DataMart, medidas que se tomaron para la limpieza de los datos y cómo se actuó ante anomalías en la información. Mencionaremos además el dataset utilizado para dicha tarea y como está estructurado.

Por último se mencionan algunas posibles mejoras a la práctica que se pueden gestionar que no se han llegado a probar por falta de tiempo, como integración de datos utilizando distintos datasets.

Sesión de prácticas I

Creación de diagrama en estrella.

Esta parte de la práctica consiste en la creación de un diagrama en estrella que representa el DataMart, para su creación se utilizó la herramienta **DBDAP** proporcionada en la documentación de la asignatura. Previamente a la realización del diseño se analizaron los requerimientos del guión de la práctica, con el fin de poder resolver cualquier consulta futura definiendo así las posibles tablas y sus atributos.

Como metodología de diseño del DataMart se siguió la metodología de Kimball detallando a continuación cada fase:

1. Proceso de negocio.
2. Selección del grano.
3. Identificación de las dimensiones.
4. Identificación de los hechos.

Las cuatro fases mencionadas anteriormente fueron necesarias para el desarrollo del diagrama en estrella que compone el DataMart generado con la herramienta DBAP. Dicha herramienta ofrece la utilidad de exportar el diagrama en formato SQL para generar la base de datos que se utiliza en la práctica 2.

Definición de las Fases de Kimball

Proceso de negocio

El proceso de negocio utilizado es el ámbito de valoración de películas, no solo por que fue la proposición del guión de prácticas, sino porque además es un tema muy utilizado con el que se suele encontrar gran cantidad de información y datos, lo cual es favorable para el desarrollo de la asignatura.

Selección del grano

Como se estudió en la asignatura el grano del DataMart debe ser lo más detallado posible, por esta razón, teniendo en cuenta que el principal análisis se realizaría sobre valoraciones de películas por usuarios, el grano fue pensado para almacenar cada valoración de cada película por cada usuario.

Identificación de las dimensiones

En esta fase se analizaron las posibles dimensiones que podrían resultar favorables para la explotación del DatMart. Dichas dimensiones consisten en tablas relacionadas directamente a la tabla de hechos, en ellas se ha almacenado toda la información que se considera relevante y atractiva para su explotación.

Se utilizó el concepto de bridges tables estudiado en clase para resolver posibles problemas lógicos en la información, como las relaciones muchos a muchos entre las tablas de dimensión y la tabla de hechos. Este paso se decidió pensando en la continuación sobre la práctica 2, ya que de no implementar las bridges tables mucha información se perdería durante el proceso ETL.

Identificación de los hechos.

El principal hecho identificado para la práctica fue el atributo valoraciones de películas, en este caso la valoración es considerado una dimensión degenerada, ya que en lugar de tener una dimensión para las valoraciones se agregó directamente a la tabla de hechos por la simplicidad y redundancia del resto de atributos.

Diseño del DataMart con la herramienta BDAP

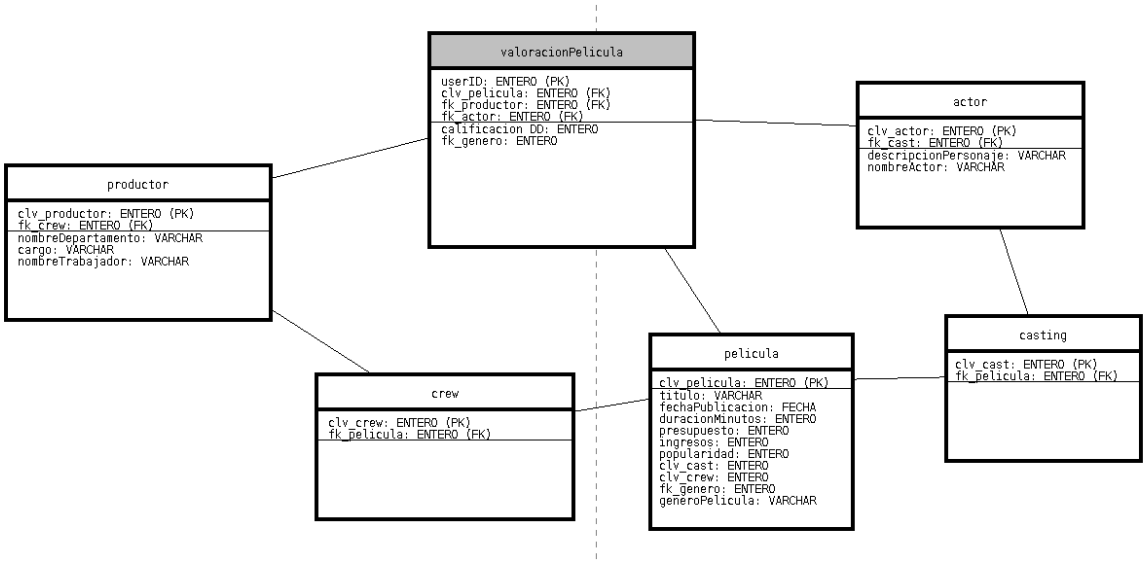

Imagen 1: Diagrama en estrella del DataMart

En la imagen 1 se observa el diagrama final correspondiente al dataMart finalizado, en el que se aprecia las siguientes tablas:

Tabla de hechos

Contiene el nivel de detalle más fino posible para la información almacenada, que representa la calificación o valoración de cada película, representando dicho atributo como un atributo **DD** de dimensión degenerada.

Tabla Película

En ella se encuentra la información genérica de cada película como:

- Título.
- Fecha de publicación.
- Duración en minutos.
- presupuesto.
- ingresos.
- popularidad.

Tablas Actor y Productor

Contienen información de las personas presentes en el rodaje de cada película como:

- Nombre de la persona.
- Cargo.
- Departamento.
- Papel interpretado.

Bridges Tables

Las bridges tables **casting** y **crew** permitieron en el DataMart la lógica necesaria para permitir insertar películas con varios actores y productores, es decir relaciones **N:M.** Esto claramente genera un impacto en el tamaño de la fact table, sin embargo con este diseño el DataMart gana capacidad de análisis al tener en una misma tabla actores, productores, películas y valoraciones en una misma tabla.

Gestor de Base de datos y gestión de archivos

Como gestor de base de datos para las prácticas se ha utilizado postgres, instalando una versión en local en la laptop de cada integrante del equipo, utilizamos el almacenamiento en drive con las cuentas de la universidad para gestionar los ficheros como creación de las tablas de la base de datos y los ficheros de población de tablas.

Limitaciones de la herramienta BDAP

Como es de esperar una integridad referencial de claves externas en una tabla de sql, requiere la primary key de la tabla a la que se esta referenciando, en este caso tanto la tabla de actores como la tabla de productores, poseen una primary key compuesta del usuario (Actor o Productor) y el id de su desempeño en el desarrollo de la película, este detalle no se ha podido representar en la herramienta ya que no se permiten agregar mas de un atributo como primary key a una tabla, por la misma razón no podemos referenciar claves externas a dichas primary key.

Solución

Para solventar este inconveniente se utilizó la herramienta de exportación SQL del dataMart y se modificaron dichas claves primarias y extranjeras en las tablas correspondientes, además se modificaron un poco los tipos de datos como longitud de almacenamiento, para mejorar la inserción de datos en la práctica 2.

Sesión de prácticas II

DataSet

Como Dataset para la sesión de práctica II se utilizó el conjunto de datos denominado The movies DataSet desde el sitio web de kaggel [1] que contiene una serie de ficheros csv con la siguiente información:

- **movies_metadata:** Archivo principal que contiene los datos de 45,000 películas, tales como título, país de producción, compañía de producción, presupuesto, recaudo, id de la película referenciado por los demás ficheros, etc.
- **credits:** contiene información del casting y crew de cada película, almacenando en cada fila el **id** de la película, y en formato json el cast y crew de cada película.
    - **cast**: contiene información como el nombre de los actores, identificadores, papel interpretado.
    - **crew**: contiene nombre del equipo de producción, cargo, departamento e identificador de cada uno.
- **Rating**: contiene las valoraciones de las películas por usuario y una valoración entre 1 y 5.

El archivo **Rating** contiene 26 millones de valoraciones, sin embargo esta cantidad se redujo durante el proceso ETL.

Estos tres ficheros son de los que se ha extraído la información almacenada en el DataMart, de **movies_metadata** se extrajo toda la información almacenada en la dimensión de películas.

Del fichero **credits** se extrajeron todos los datos relacionados con actuaciones y productores de cada película almacenada en la dimensión películas, toda esta información estaba almacenada en matrices JSON, definiendo en cada celda un matriz correspondiente a cada película.

Por último del fichero **Rating** se extrajeron las valoraciones de películas por usuarios.

Proceso ETL

Como parte del proceso **ETL** el primer reto fue entender la estructura de los datos del dataset, ya que, como se mencionó anteriormente, cada el fichero credits contiene una serie de matrices json para cada película, por lo tanto se tuvo que validar primeramente que estuvieran directamente relacionados con el resto de ficheros, es decir, se verificó la **veracidad de los datos**.

El segundo reto fue encontrar la manera adecuada para poblar los hechos y dimensiones del DataMart partiendo de los ficheros csv. Debido a la complejidad de esta tarea se tomó la decisión de escribir un programa utilizando java como lenguaje de desarrollo para tratar los datos.

Limpieza de datos

Anterior al procesamiento de datos se realizó limpieza genérica sobre los ficheros csv utilizando filtros sobre los campos de los archivos, los errores solucionados de esta manera fueron:

- **Películas inexistentes.**

Antes de empezar la lectura de los datos con el programa desarrollado fue necesario realizar algunas tareas de limpieza en los csv, por ejemplo, en el fichero de película existían películas sin ninguna información más que solo su **ID**, desde luego esta información no nos sería útil, por lo que se procedió a su eliminación de este y de los demás ficheros, esto por supuesto tuvo una cantidad mayor de datos afectados en la tabla de valoraciones.

- **Espacios Vacíos**

Al igual que identificamos películas inexistentes filtramos aquellos campos que no tuviesen información, en algunos casos si dicho campo afectaba pocas tuplas eliminamos la tupla de la data, sin embargo hubieron algunos datos que afectan en grandes cantidades la inserción de datos del DataMart, en este caso se decidió no utilizar este campo para no afectar el volumen de datos.

Además de estas medidas escribimos en en el programa en java una serie de excepciones que permitieran escribir la línea que causaba el error en la lectura de datos.

Procedimiento de carga

El proceso de carga de datos consistió en la creación de archivos **SQL** que contienen los insert generados para la población de las tablas, estos archivos fueron ejecutados uno a uno posteriormente mediante la línea de comandos de postgres. Dichos archivos SQL fueron creados utilizando un programa hecho en java escrito por los integrantes del equipo de prácticas, un programa básico que lee cada línea de los ficheros **csv** que ya fueron tratados como se describe anteriormente. Los datos interesantes que se cargaron son los que planteamos en la primera práctica, puede verse el código tanto del programa java como de los ficheros generados por el mismo y el script SQL que conforma la base de datos en el siguiente enlace **[2].**

El procedimiento descrito en el párrafo anterior para la población de la base de datos no aplica para la inserción de las relaciones de actores y productores presentes en la tabla de hechos, dichos campos fueron generados y cargados modificando las tablas y generando consultas SQL para que se insertarán en ellos información que había sido guardada a partir de otras tablas.

A mejorar

Como continuación de la práctica se podría poblar un poco más el DataMart utilizando diferentes conjuntos de datos, para esto haría falta analizar si el conjunto de datos es compatible con la actual estructura del DataMart o modificar el mismo para obtener una mayor variedad de datos.

También se podrían generar nuevas cargas de datos sobre un nuevo DataMart que tenga como punto de partida la actual base de datos, esto con el fin de tener una estructura de montaje más limpia, evitando así la población de columnas en la tabla de hechos mediante alteraciones de columnas e inserts generados a partir de consultas generadas sobre tablas previamente pobladas.

Bibliografía

1. Conjunto de datos utilizado en la práctica II:

   ttps://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

2. Código desarrollado para la carga del dataset:
   https://github.com/794866/BIGDATA