Reporte Proyecto Final

Almacenes y Minería de Datos, 2020-2
Facultad de Ciencias, UNAM

Gómez Mora Héctor Eduardo **(312296414)**

312296414@ciencias.unam.mx

Deposición de Datos: Las datos así como la implementación del cubo pueden encontrarse en el siguiente repositorio de GitHub https://github.com/hemora/amd-proyecto.

1. Introducción

Dada la proliferación de la industria de los videojuegos y del incremento en la competitividad que esto conlleva, se ha vuelto necesario el uso de herramientas y técnicas que permitan realizar un análisis del mercado a través de los datos que de las ventas puedan obtenerse.

Sin embargo, para ello es necesario disponer de una infraestructura apta para la consulta de grandes cantidades de datos asociados de forma compleja. Un esquema que permita asociar jerárquicamente a los datos con la finalidad de observar diversos hechos a través de varios gradientes de precisión.

Ante estas disyuntivas es que se propone el uso de un Data Mart (DM) para realizar esta tarea.

2. Objetivo

El presente trabajo busca reportar el proceso de construcción de un DM a partir de un conjunto de registros concernientes a la venta de videojuegos en distintas partes del mundo y en distintos años. Esto con la finalidad de facilitar y automatizar la obtención de información valiosa con capacidad de otorgar ventaja competitiva; además de mostrar algunos de los resultados obtenidos de su explotación.

3. Metodología

3.1. Obtención e Inspección de Datos

La obtención de los datos se realizó a través del sitio Kaggle de donde se logro extraer un dataset [1] de 16598 registros dedicado a almacenar la cantidad de ventas realizadas para distintos videojuegos en distintas regiones del mundo desde 1980 hasta 2020. Este se compone de los siguientes atributos:

Nombre	Descripción
Rank	El ranking sobre las ventas registradas.
Name	El nombre del videojuego
Platform	El nombre de la plataforma de lanzamiento.
Year	Año de lanzamiento
Genre	Categoría de género
Publisher	Empresa distribuidora del juego
NA_sales	Ventas en Norte América (en millones)
EU_sales	Ventas de Europa (en millones)
JP_sales	Ventas en Japón (en millones)
Other_sales	Ventas en las regiones restantes (en millones)
Global_sales	Ventas a nivel global (en millones)

Cuadro 1: Atributos del Data Set

3.2. Creación de Esquema Multidimensional

La principal temática identificada fue la cantidad de ventas sucedidas para cada videojuego. En particular, la cantidad de ventas globales representará la medida principal sobre la cuál formar el esquema del Data Mart. Dado que los demás atributos concernientes a ventas pueden ser agregados para obtener la cantidad de ventas totales se decidió en un principio separarlos como una dimensión separada.

Puesto que, para los atributos restantes no se logró idear una jerarquía natural que los asociara de forma idónea se decidió tratarlos como dimensiones independientes con excepción de **Rank**, **Name**, **Year** que figuran como atributos de la tabla de hechos.

Para los atributos restantes, con la finalidad de agregar dimensiones y enriquecer la potencial información que pueda obtenerse de la explotación del DM, se realizó un preprocesamiento para agregar atributos y construir dimensiones. Para ello se propusieron los siguientes cambios:

Platform

Con respecto al avance tecnológico que han atravesado las consolas y servicios de videojuegos se ha vuelto común clasificarlas bajo la *generación* a la que pertenecen. De lo cuál un nuevo atributo `generation` será el encargado de relacionar a cada plataforma presente en nuestra base de datos con alguna de las 8 generaciones trascurridas hasta ahora.

Genre

Por su parte, la clasificación PEGI (Pan European Game Information) es un elemento indispensable durante el proceso de promoción, venta y distribución de un videojuego ya que determina en gran parte el sector de personas que lo adquieren. Para modelar este hecho un nuevo atributo `pegi` será el encargado de relacionar a cada uno de los géneros existentes en nuestra base de datos con un rating que varía entre 3,4,6,7,12,16 y 18.

Publisher

Para las empresas productoras y distribuidoras de videojuegos se propone enriquecer su dimension al agregar los atributos `country` y `region` que otorgan información acerca de la ventaja que poseen ciertas regiones y países con respecto a la venta y distribución de videojuegos, y contrastar con el impacto que tienen dentro de sus propios países/regiones.

Dada la obvia dificultad que implica incluir estos atributos manualmente y de crear registros coherentes mediante la automatización, debe resaltarse que la asignación de valores se realizó de forma **aleatoria** por lo que si bien la información obtenida de explotar el DM puede no concordar con la realidad se espera que la escencia de la consulta sea de potencial valor.

Una primer propuesta de jerarquías y esquema multidimensional de estrella se muestra acontinuación:

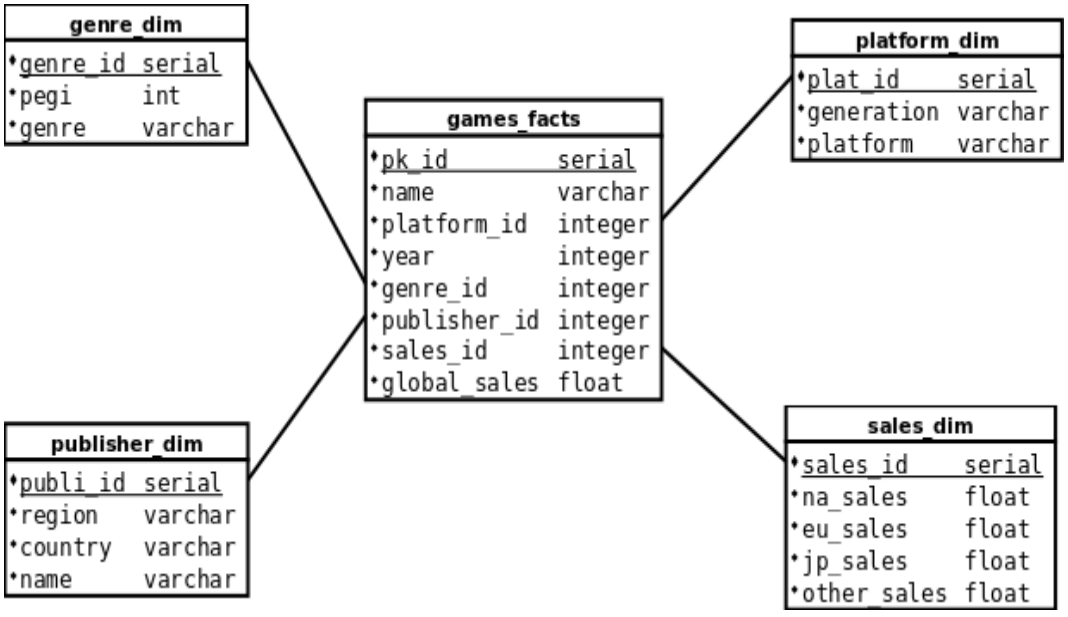

Figura 1: Primer esquema propuesto

La transformación del dataset para agregar los atributos anteriormente mencionados se realizó mediante un preprocesamiento con Python, mientras que la carga de estos últimos junto con el dataset original se realizó utilizando el DBMS PostgreSQL.

Por su parte, la implementación del esquema propuesto se realizó utilizando el software de creación de cubos **Schema Workbench** dentro de la Pentaho Platform dispuesta por Hitachi Vantara [2] a partir de los datos anteriormente cargados estableciendo como

medida al atributo `global_sales` que representa la cantidad de ventas totales obtenidas por un videojuego.

Una vez creado este último se realizaron diversas consultas MDX con la finalidad de revisar su correcta implementación y la lógica de los resultados obtenidos. A partir de lo cuál lograron encontrarse inconsistencias en el esquema.

En particular, de consultar la dimensión `sales_dim` fueron obtenidas la cantidad de ventas totales para una venta concreta correspondiente a alguna región. Esto último resulta claro que no representa una consulta o enunciado coherente en su contexto por lo que una reestructuración de las dimensiones fue necesaria.

De un segundo análisis se determinó que en lugar de una jerarquía, los atributos **NA_sales**, **EU_sales**, **JP_sales** y **Other_sales** se corresponden de mejor manera como medidas adicionales a `global_sales` y que permiten un análisis más eficiente de las ventas al poder seleccionar el mercado de interés. Como parte de estos cambios, la dimension `sales_dim` fue eliminada y sus atributos devueltos a la tabla de hechos.

A su vez, el esquema anteriormente planteado no permite presentar los nombres de los videojuegos ni su año de lanzamiento, información que podría llegar a ser de interés, puesto que dichos atributos figuran en la tabla de hechos. De esto último, fue necesario añadir una nueva dimensión `games_dim` que incluyese a dichos atributos y los ligara con la tabla de hechos mediante una llave subrogada.

El esquema actualizado se presenta acontinuación:

4. Resultados

Del DM resultante, se presentan a continuación algunas aplicaciones que se consideran interesantes y que pueden llegar a traducirse en ventaja competitiva:

Mediante manipulación de expresiones multidimensionales puede encontrarse el número de ventas sucedidas en Norte América para todos los géneros de videojuegos registrados:

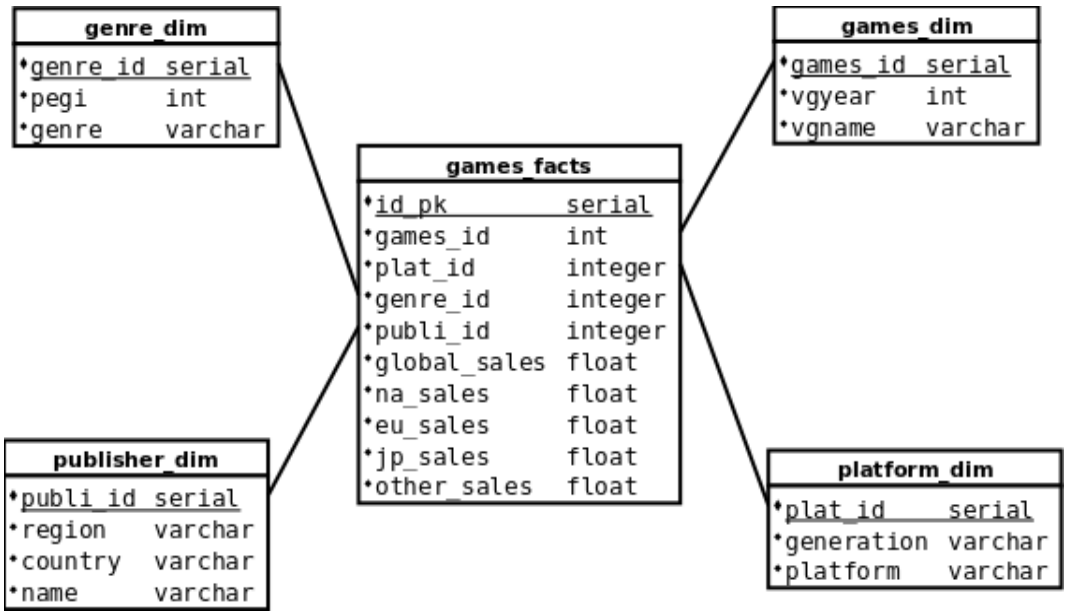

Figura 2: Segundo esquema propuesto

```
SELECT
  [Genre].[Genre].members ON COLUMNS,
  [Measures].[NAUnit]ON ROWS
FROM [VG Sales]
```

De ello, se puede notar como **Puzzle** figura como el género con menos videojuegos vendidos en Norte América con tan sólo $126.43 millones obtenidos.

Una vez identificada esa tendencia en el mercado Norte Americano, puede contrastarse con respecto a los demás:

```
SELECT
  [Genre].[Genre].[Puzzle] ON COLUMNS,
  {[Measures].[NAUnit], [Measures].[JPUnit], [Measures].[EUUnit]}
  ON ROWS
FROM [VG Sales]
```

De lo cuál, puede observarse como dicha no cambia en los restantes dos mercados de importantancia sino que resulta más notoria, recaudandose tan sólo $45.84 y $55.58 millones en los mercados Europeo y Japonés respectivamente.

A partir de esto último un potencial usuario puede encontrar pertinente no invertir en la producción de videojuegos de este género o estudiar los intereses de este nicho para poder explotarlo dada la baja competitividad del mismo.

Otro cas en el que el uso del cubo resulta útil es cuando se desea conocer el impacto de un producto o una industria en un mercado específico. Véase por ejemplo,

```
SELECT
  CROSSJOIN (
    [Publisher].[Region].[Asia],
    [Platform].[Platform].[GB]
  ) ON COLUMNS,
  {[Measures].[NAUnit], [Measures].[EUUnit], [Measures].[JPUnit]} ON ROWS
FROM [VG Sales]
```

que obtiene el total de ventas debidas a títulos pertenecientes a distribuidoras asiáticas desarrollados para la consola Game Boy con una perspectiva de los tres mercados más significativos, lo cuál nos permite hacer una análisis comparativo.

De ello puede observarse como las ventas para esta plataforma son más prolíficas en el mercado norteaméricano llegando a representar el 59 % de las ventas analizadas. El procentaje restante se divide en 12 % Y 27 % para los mercados europeo y japonés respectivamente.

A partir de esta información el usuario puede decidir si encaminar sus esfuerzos de marketing para un determinado producto en cierto mercado o mejorar la gestión e infraestructura dentro de su propio contexto.

Por otro lado, gracias a la capacidad que tiene el modelo multidimensional de presentar información con niveles de granularidad distintos se puede realizar un análisis similar al anterior realizando la operación drilldown sobre la dimensión `publisher_dim` de forma que,

```
SELECT
  CROSSJOIN (
    [Publisher].[Publisher Name].[Bethesda Softworks],
    [Platform].[Platform].[X360]
  ) ON COLUMNS,
  {[Measures].[NAUnit], [Measures].[EUUnit], [Measures].[JPUnit]} ON ROWS
FROM [VG Sales]
```

representa la cantidad recaudada por las ventas de títulos o servicios desarrollados por *Bethesda Softworks* para la consola Xbox360, de lo cuál puede observarse como el 65 % de la recaudación de sus productos es debido a su propio mercado de origen, siendo los restantes 29 % y 5.29 % debidos a las recaudaciones por parte del mercado europeo y japonés respectivamente.

De esta información el usuario puede decidir si limitar sus esfuerzos de maketing a su

propia región de origen o en cambio, establecer relaciones con distribuidoras asiáticas para promover sus productos a través de las plataformas predominantes en esos mercados.

Referencias

[1] Video Game Sales [Data Set] Disponible en:
https://www.kaggle.com/gregorut/videogamesales

[2] Pentaho Schema Workbench [Software] Disponible en:
https://help.pentaho.com/Documentation/8.3/Products/Pentaho_Schema_Workbench