Modelando OLAP

Autor
Martín Vladimir Alonso Sierra Galvis

Gestión de Datos, Maestría en Ciencia de Datos
Pontificia Universidad Javeriana Cali

Versión 2.0
Santiago de Cali, marzo de 2023

Tabla de contenido

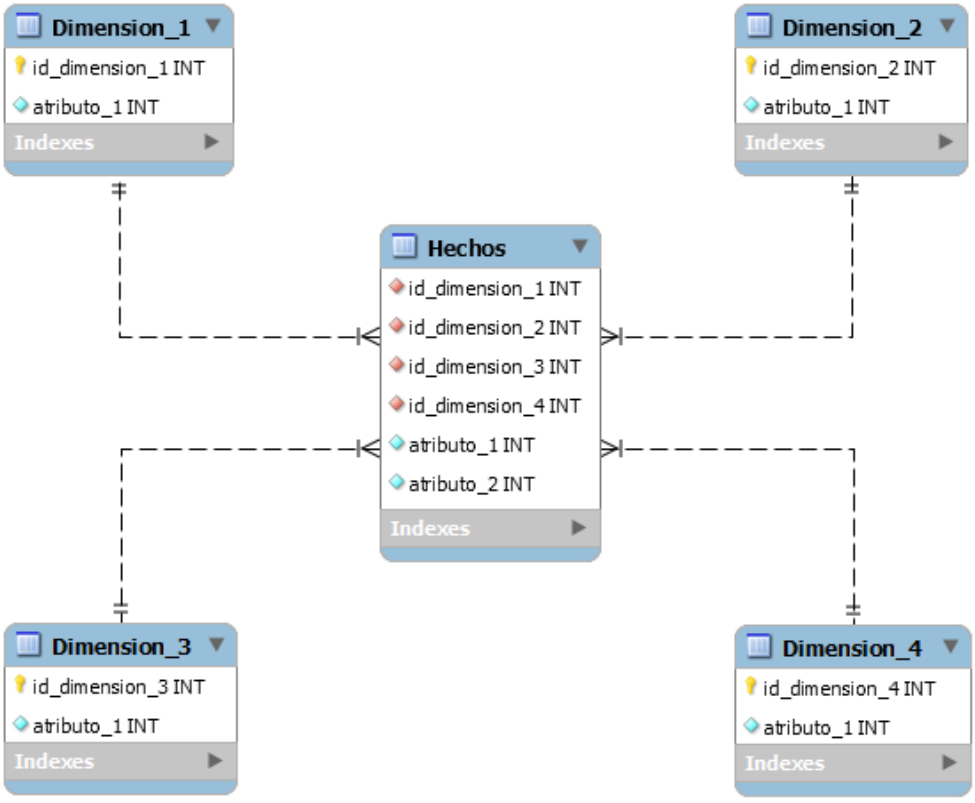
Modelo de una Bases de Datos OLAP

Consideraciones

En la segunda unidad del primer módulo, llamada Bases de Datos Relacionales, vimos el concepto de modelo y cómo el Modelo Relacional era una herramienta fundamental en la actividades de Gestión de Datos. Esto tanto para profesionales con cierta autoridad sobre los datos como los Ingenieros de Software o los Ingenieros de Datos, como para los Científicos de Datos, cuyo rol principal es acceder a los datos para realizar diversos procesos con ellos. Para el tema que nos atañe en esta unidad, la historia será similar. Al utilizar el concepto de tablas y relaciones, la implementación de una Base de Datos OLAP implica, en primera instancia, tomar un Modelo Relacional y, a partir de unos requerimientos de análisis, implementar un Modelo Multidimensional equivalente. El Modelo Multidimensional resultante nos servirá para realizar diversas actividades de gestión como implementación de un Data Mart físico; alimentación del Data Mart mediante procesos ETL, usando los datos de la base de datos relacional original; y acceder a los datos mediante un visualizador de Cubos OLAP. En esta lectura aprenderemos cómo construir un Modelo Multidimensional de Estrella, el modelo más utilizado para la implementación de Bases de Datos OLAP.

El Modelo Estrella

El Modelo Estrella es el modelo más simple y, a la vez, el más utilizado entre los Modelos Multidimensionales, y sirve para consolidar una gran cantidad de datos en un Data Mart. Se llama así porque su diagrama se parece a una estrella. Como pasaba con el Modelo Relacional, el Modelo Estrella adopta el concepto de tablas y de relaciones entre tablas para su estructura básica. La diferencia entre los dos modelos radica en que el Modelo Relacional, en un principio, no hace ninguna distinción entre las tablas que conforman el modelo, mientras que al trabajar con un Modelo Estrella podemos hablar de dos tipos de tablas diferentes: las tablas que representan **Hechos** y las tablas que representan **Dimensiones**. La siguiente imagen muestra el diagrama general del modelo, diagrama que ya habíamos visto anteriormente en la lectura **Conceptos Básicos**, insumo previo de esta unidad.



En el diagrama podemos identificar los componentes principales: en el centro podemos ver la tabla de Hechos y a su alrededor, a modo de satélites, se encuentran las tablas de Dimensiones. Se dice que la tabla central de Hechos se encuentra en su Tercera Forma Normal o 3FN, esto debido a que, si nos fijamos bien en los campos que la conforman, nos daremos cuenta que los datos que almacena son valores unitarios y la mayoría representan Llaves Foráneas que se relacionan con las Llaves Primarias de cada una de la Dimensiones. Por otro lado, estas tablas de Dimensiones no están normalizadas pues, como veremos en los videos, se pueden presentar casos en que los datos almacenados en una Dimensión se repitan. Lo anterior implica que, contrario a lo que pasaba con las Bases de Datos Relacionales, los Data Mart permiten almacenar datos redundantes, traduciéndose esto en una mejora de rendimiento al momento de realizar consultas complejas enfocadas al análisis de datos. La razón: cuando trabajamos con una Base de Datos Relacional necesitamos hacer muchas consultas cruzadas de datos debido, principalmente, al uso de las estructuras normalizadas para evitar la redundancia de datos; por lo descrito anteriormente, esto no pasará en un Data Mart.

Pero, cuando hablamos de tablas de Hechos y tablas de Dimensiones, ¿a qué nos referimos? Para entender la razón de ser de cada uno de estos componentes, debemos tener claro lo siguiente:

1. Los **Hechos** representan los procesos de negocio de la organización o del stakeholder que se encuentre interesado en realizar el análisis de sus datos. El ejemplo más común es el proceso de **ventas**. Además de las Llaves Foráneas que enlazan con las tablas de Dimensiones, la tabla de Hechos almacena métricas que corresponden a hechos numéricos. Si estamos representando el proceso de ventas, una posible métrica puede ser la **cantidad de productos vendidos**.
2. Las **Dimensiones** representan determinados puntos de vista desde el que se puede analizar un proceso de negocio. Por ejemplo, siguiendo con el proceso de ventas, un punto de vista que podríamos utilizar para analizar dicho proceso son los productos en sí. Debemos tener presente que los datos almacenados en las dimensiones son de tipo descriptivo.

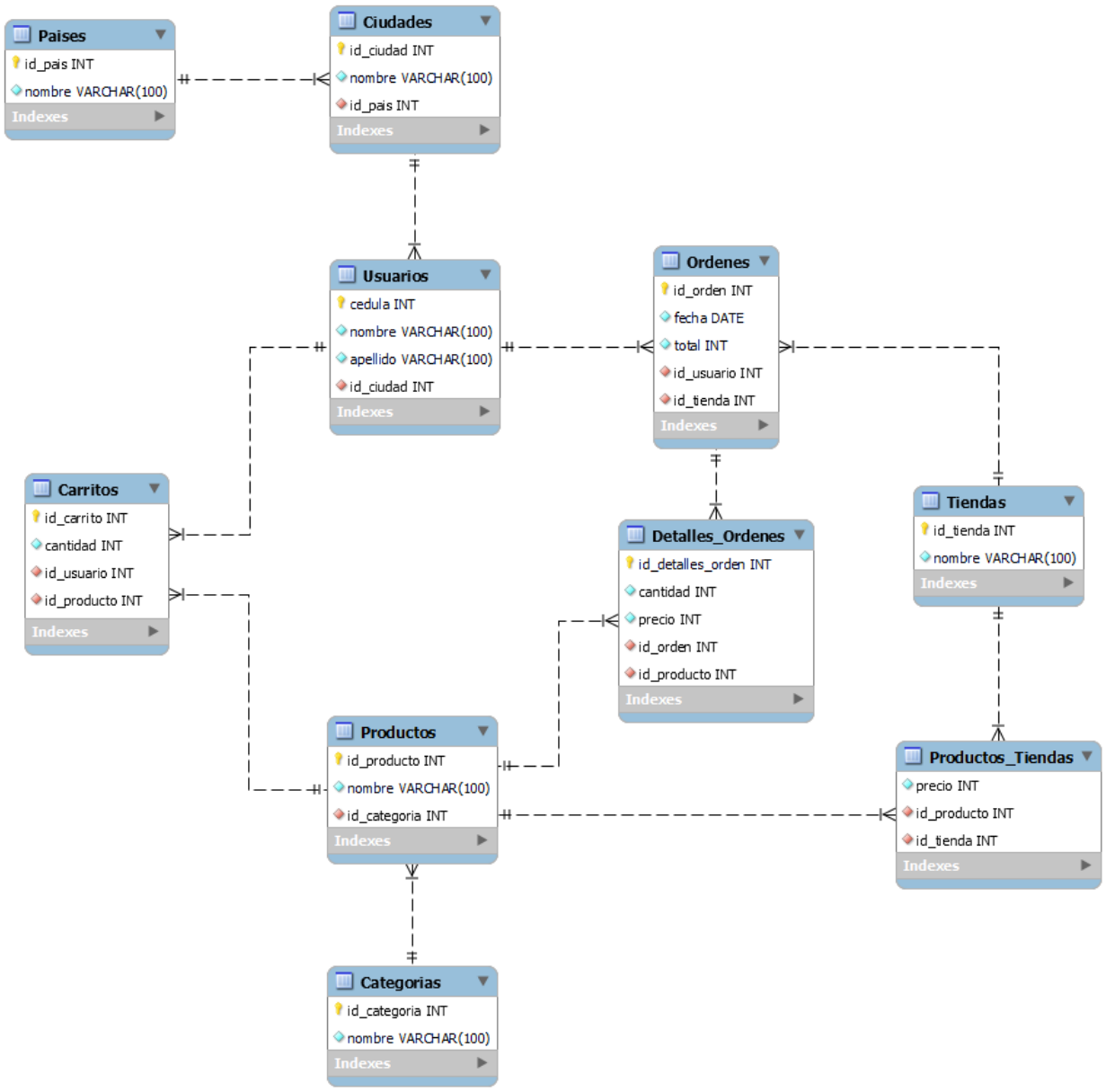
Veamos ahora cuáles son los pasos a seguir para poder construir un Modelo Estrella a partir de un Modelo Relacional.

Proceso para construir el Modelo Estrella

Para realizar los pasos de construcción del Modelo Estrella, vamos a tomar como ejemplo un Modelo Relacional que representa una Base de Datos Relacional encargada de almacenar los datos de una super app en la que los usuarios pueden comprar diversos productos de diferentes tiendas.

Entendiendo el Modelo Relacional original

El Modelo Relacional de la base de datos de la super app lo podemos ver en la siguiente imagen.



Del Modelo Relacional de ejemplo podemos interpretar, entre otras cosas, lo siguiente:

1. Las ciudades a las que pertenecen los usuarios registrados en la base de datos. La relación indica que un usuario pertenece a una única ciudad y que una ciudad puede tener muchos usuarios asociados.
2. Los países a los que pertenece cada una de las ciudades registradas en la base de datos. La relación indica que una ciudad pertenece a un único país mientras que un país puede tener muchas ciudades.

3. Los productos que tiene un usuario en su carrito de compras en determinado momento. La tabla Carritos es una tabla intermedia que indica que un usuario puede tener muchos productos en su carrito y un producto puede estar en muchos carritos de usuarios al tiempo.
4. Las órdenes de compra que generan los usuarios al comprar productos en la super app. La relación indica que un usuario puede generar muchas órdenes de compra y que una orden de compra pertenece únicamente a un usuario determinado. En este punto es interesante notar que la tabla Ordenes actúa también como una tabla intermedia donde podemos apreciar que un usuario puede generar sus órdenes de compra en diferentes tiendas y una tienda puede generar órdenes de compra de muchos usuarios.
5. Los productos que se encuentran registrados en una orden de compra. La tabla Detalle_Ordenes es una tabla intermedia que indica que una órden de compra puede tener muchos productos registrados y que un producto puede, a su vez, registrarse en muchas órdenes de compra.
6. Las categorías a las que pertenecen los productos. Para el caso del ejemplo se asume que un producto pertenece a una única categoría mientras que una categoría puede contener varios productos.
7. Las tiendas donde se encuentran los productos que se ofrecen en la aplicación. La tabla intermedia Productos_Tiendas indica que un producto se pueden encontrar en diversas tiendas y una tienda puede tener muchos productos.

Finalmente, a modo de aclaración y por simplicidad, vamos a suponer que las órdenes de compras representan las ventas realizadas.

Los requerimientos de análisis

Al comenzar a ejecutar cada uno de los pasos para construir el Modelo Estrella a partir del Modelo Relacional, es importante saber que el primero sigue un proceso muy diferente a este último debido, principalmente, a que un Data Mart es el resultado de ciertos requerimientos de análisis muy específicos. Los requerimientos de análisis dependen del análisis que la organización o stakeholder quiera realizar sobre los procesos de su negocio. Esta es la razón por la que en un Data Warehouse podemos encontrar muchos Data Marts almacenados, mismos que, a su vez, pueden ser la implementación física de diferentes Modelos Multidimensionales. Lo interesante es que, posiblemente, dichos Modelos Multidimensionales se

hayan construido a partir de un único Modelo Relacional. Vamos entonces a suponer que nuestros requerimientos de análisis son:

1. Total de productos tecnológicos vendidos para la ciudad de Santiago de Cali en el año 2019.
2. Total de ventas por categoría y por tienda en el año 2018.

Análisis de los hechos

El primer paso a seguir es analizar los **hechos**. Recordemos que los hechos hacen referencia a un proceso de negocio determinado que una organización o stakeholder esté interesado en analizar. Tomando los requerimientos de análisis del ejemplo, podemos identificar que el proceso de análisis que el stakeholder desea realizar está relacionado con las ventas de los productos, por lo que el hecho en este caso es **productos vendidos**.

Análisis de las dimensiones

Una vez se ha analizado el hecho, debemos continuar con el análisis de las **dimensiones**. Como vimos anteriormente en este mismo documento, las dimensiones son puntos de vista desde los que podemos analizar el hecho identificado. Para esto, debemos revisar los requerimientos de análisis del ejemplo. En el primer requerimiento podemos notar que se desea saber el total de productos **tecnológicos** vendidos. La palabra, tecnológicos, hace referencia a una categoría de producto. Por lo tanto, **categoría** podría ser un punto de vista. Así mismo, según el requerimiento, este total de productos vendidos se desea calcular para una ciudad y un año particular que son Santiago de Cali y 2019 respectivamente. Aquí podemos, entonces, identificar otros dos puntos de vista: **ciudad** y **fecha**. Si realizamos este mismo análisis para el segundo requerimiento podemos obtener los puntos de vista **categoría**, **tienda** y **fecha**. Así pues, las dimensiones que debemos modelar para el ejemplo son categoría, ciudad, fecha y tienda.

Análisis de los indicadores

El tercer paso consiste en analizar los indicadores. Los indicadores son métricas con las que se desea medir los hechos o procesos de negocio. Es por esta razón que a los indicadores, comúnmente, se los conoce con el nombre de **medidas**. Si leemos detenidamente los requerimientos de análisis podremos ver que el hecho **productos vendidos** se quiere analizar

mediante dos medidas principales: el total de productos vendidos y el total de ventas, entendiéndose esta última como la cantidad de dinero recaudada por la venta de los productos. Estos serían los indicadores a tener en cuenta en nuestro modelo.

Análisis de las jerarquías

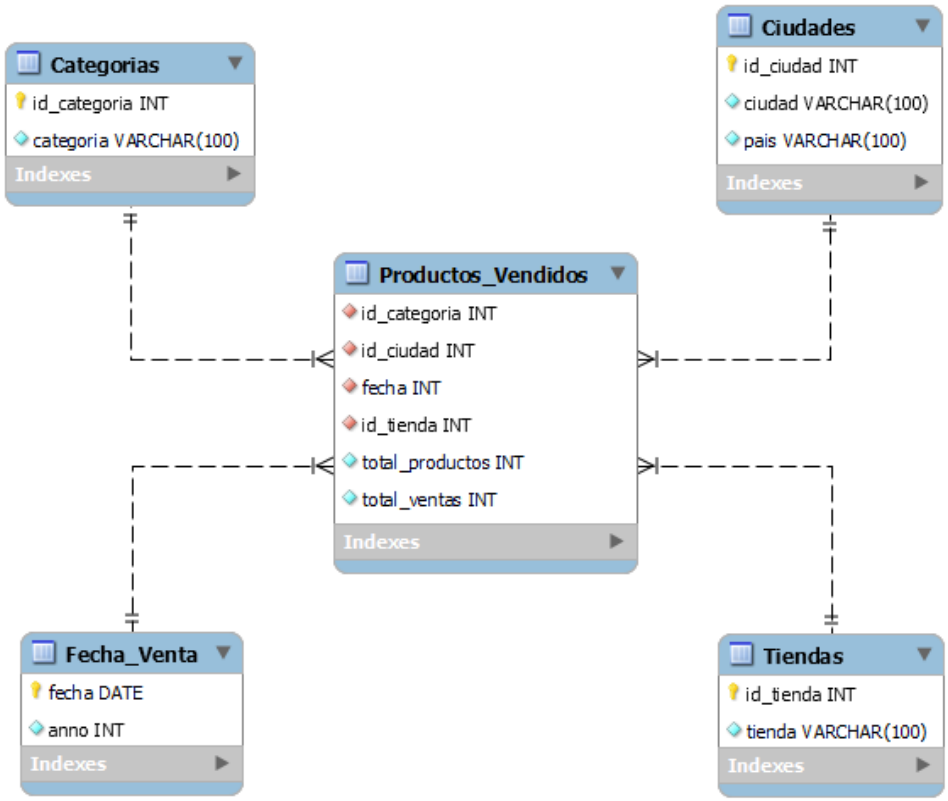
Como cuarto y último paso de análisis tenemos el análisis de las jerarquías, mismo que se realiza sobre las dimensiones que identificamos anteriormente. Para realizar este paso, debemos tener presente que toda tabla dimensión en el modelo tiene dos elementos principales: una llave primaria y un **nivel**. Este último contiene el dato descriptivo principal de la dimensión, siendo representado, normalmente, por un nombre. Ahora bien, puede ser que de un nivel se desprenda un segundo nivel, del segundo un tercero y así sucesivamente. Esto es lo que se conoce como una **jerarquía** de niveles. Por convención de las estructuras multidimensionales, las jerarquías en una dimensión cuentan con, mínimo, un nivel, mismo que nombramos unas líneas más arriba. Identificar y modelar jerarquías de diferentes niveles puede ser muy ventajoso para resolver requerimientos de análisis complejos o para expandir a futuro las preguntas de análisis. Por ejemplo, previamente identificamos que, para responder a nuestros requerimientos de análisis, necesitábamos una dimensión de fecha. La cuestión es la siguiente: si somos más específicos, de esa fecha necesitamos únicamente el año. La buena noticia es que a partir del mismo dato fecha podemos obtener el año, lo que significa que aquí podemos modelar una jerarquía. Ahora, supongamos que, para nuestro ejemplo, quisiéramos saber las ventas de determinado producto tanto por ciudad como por país. Pasa exactamente lo mismo que con la fecha: a partir de la ciudad podemos obtener el país. ¿Pero cómo implementar esta última jerarquía? Esto lo podemos lograr gracias a la organización de los datos en la Base de Datos Relacional, misma que podemos ver en el Modelo Relacional original. La ciudad está relacionada con el país y, si pensamos en el orden de las cosas en la realidad, una ciudad depende o pertenece directamente a un país. Por lo tanto, para nuestro ejemplo, podríamos modelar las siguientes jerarquías.

1. Categoría.
2. Ciudad → País.
3. Fecha → Año.
4. Tienda.

Diagrama del Modelo

Ya identificados hechos, dimensiones, indicadores o métricas y jerarquías, podemos proceder a realizar el diagrama del Modelo Estrella. Para construir el diagrama es importante considerar que:

1. El **hecho**, en este caso **productos vendidos**, se transforma en la tabla central del modelo.
2. Las **dimensiones**, en este caso **categoría**, **ciudad**, **fecha** y **tienda** se transforman en las tablas satélites alrededor de la tabla de hechos.
3. Los **indicadores**, en este caso **total productos** y **total ventas**, pasan a convertirse en los atributos de la tabla de hechos.
4. Las **jerarquías de niveles** pasan a conformar los atributos de las tablas de dimensiones.
5. Las tablas de dimensiones contienen Llaves Primarias que son referenciadas como Llaves Foráneas en la tabla de hechos.



Realizando los reemplazos correspondientes al diagrama general del Modelo Estrella, el diagrama final del ejemplo que hemos venido trabajando en esta

lectura quedaría como se muestra en la imagen anterior. Recordemos que este diagrama es un resultado específico para los requerimientos de análisis del ejemplo mencionado. Esto quiere decir que, dependiendo de dichos requerimientos, podríamos obtener un Modelo Estrella diferente. Comparándolo con el Modelo Relacional original podemos destacar que:

1. En la dimensión Ciudades existe redundancia de datos pues almacena los nombres de las ciudades y los nombres de los países a los que pertenecen dichas ciudades. Como consecuencia de esta organización, pueden registrarse una gran cantidad de nombres de países repetidos. Como ya se ha mencionado, este comportamiento en los datos es una buena señal en un Modelo Estrella ya que permite consultas más eficientes en cuanto a rendimiento.
2. Conceptualmente hablando hay un cambio muy notorio: en un Modelo Relacional, lo normal es modelar los datos relacionados a fechas como un atributo o campo de alguna tabla. Aquí, en el Modelo Estrella, la fecha se modela como una tabla de nombre Fecha_Venta. Esto se debe a que, en este modelo, el concepto correcto es dimensión y no tabla.

Con esto hemos terminado todo el proceso para construir el Modelo Estrella a partir de un Modelo Relacional y unos requerimientos de análisis. Este Modelo nos servirá para organizar los datos en una estructura que nos permitirá realizar consultas más eficientes para el análisis de datos: el Data Mart. En este punto puede surgir en nosotros la siguiente duda: ¿qué herramientas debemos utilizar para implementar y trabajar físicamente un Data Mart? En un principio, podemos implementarlo utilizando un Gestor de Bases de Datos Relacionales como Oracle Database. Esto debido a que las estructuras, tanto de hechos como de dimensiones, corresponden a tablas que se relacionan mediante llaves primarias y foráneas. Posteriormente, utilizaremos esta estructura física del Data Mart para construir Cubos OLAP y como fuente destino de datos de procesos ETL, usando para ello herramientas especializadas como Schema Workbench y Data-Integration, respectivamente. Tanto el Data Mart como el Cubo OLAP serán elementos muy importantes para visualizar datos desde una estructura multidimensional. Todo esto corresponde a los temas principales que abordaremos en los videos de esta unidad.

Referencias

1. Auribox Training. *Modelo Dimensional OLAP para la elaboración de un Data Mart*. Junio de 2017. [En línea]: disponible en https://www.youtube.com/watch?v=x2keCD2lCuk. [Accedido en 2020].
2. Auribox Training. *Tablas de hechos y dimensiones | Modelo Estrella [Data Mart]*. Junio de 2017. [En línea]: disponible en https://www.youtube.com/watch?v=HvzO18fUjqY. [Accedido en 2020].
3. Datawarehouse4u. *Star schema*. [En línea]: disponible en https://www.datawarehouse4u.info/Data-warehouse-schema-architecturestar-schema.html. [Accedido en 2020].
4. Encalada, E. - UTPL. *Bases de datos: Diseño de un Cubo OLAP (Ejemplo 1)*. Julio de 2016. [En línea]: disponible en https://www.youtube.com/watch?v=jJG0INtiOa8. [Accedido en 2020].
5. Encalada, E. - UTPL. *Bases de datos: Diseño de un Cubo OLAP (Ejemplo 2)*. Julio de 2016. [En línea]: disponible en https://www.youtube.com/watch?v=vGYCo59QNQQ. [Accedido en 2020].