IMF Business School

Universidad Camilo José Cela

Máster en Big Data & Business Analytics

Módulo V: Inteligencia de Negocio y Visualización

Documentación del caso práctico

31 de Octubre del 2018

Autor
José María Sánchez Salas
josemaria.sanchezsalas@gmail.com

Índice

Módulo V: Inteligencia de Negocio y Visualización

1 Introducción

Este documento contiene la explicación de mi solución propuesta al caso práctico del Módulo V: Inteligencia de Negocio y Visualización, del M. Big Data & Business Analytics impartido por IMF Business School. El documento contiene tres secciones principales: en la primera se muestra el enunciado el caso práctico, en la segunda se muestra la solución propuesta a todas las preguntas planteadas y por último, la tercera sección muestra las conclusiones a las que se ha llegado en la realización de este caso práctico.

2 Enunciado

El departamento antifraude de una compañía de Mystery Shopping desea hacer un seguimiento y analizar la información relativa a las encuestas que realiza en los distintos centros de sus clientes. Para ello, el cliente solicita:

- Un análisis y diseño del Data Warehouse que daría respuesta a los usuarios analíticos del departamento antifraude, suponiendo que los usuarios aún no tienen claro el tipo de análisis que quieren realizar.
- Partiendo del análisis y diseño previo realizado y usando Pentaho Data Integration, se debe realizar la implementación del proceso ETL con el objetivo de:
  - Identificar y extraer los datos de las fuentes.
  - Procesar los datos y aplicar procesos de limpieza y calidad del dato.
  - Generar y cargar los datos en el modelo físico de estrella identificado en la fase de diseño.
- Posteriormente, partiendo del análisis y diseño previo realizado y conociendo ya la tecnología seleccionada, en este caso Pentaho Business Analytics, ha de realizarse una implementación ágil del modelo multidimensional.

El objetivo en este caso es la implementación del modelo multidimensional sobre diseño del Data Warehouse que daría respuesta a los usuarios analíticos del departamento antifraude, suponiendo que los usuarios aún no tienen claro el tipo de análisis que quieren realizar.

3 Solución

3.1 Análisis de fuentes

En esta sección se va a realizar el análisis de las fuentes de las que se dispone en este caso práctico. La sección está dividida en dos secciones: la primera de ellas se hace una descripción global de las fuentes mientras que en la segunda se hace una descripción detallada de los campos que contienen cada una de las fuentes.

3.1.1 Descripción global de las fuentes

Para este problema se nos presenta una única fuente: un fichero en formato CSV que contiene la información de todos las encuestas registradas por la compañía Mystery Shopping. Este fichero CSV contiene los siguientes campos:

3.1.2 Descripción en detalle de cada campo

- `COD_LOC`. Se trata del código de centro en el que se realizó la encuesta. Es un campo alfanumérico y único que identifica al centro.
- `NOMBRE_LOC`. Se trata del nombre del centro al que corresponde el campo anterior. Es un campo alfanumérico.
- `CP`. Corresponde con el código postal del centro y es un campo numérico.
- `POBLACION`. Corresponde con la población asociada al código postal y en el que se ubica el centro. Se trata de un campo alfanumérico.
- `OFICINA`. Se trata de la oficina a la que está asignada el centro. Se trata de un campo alfanumérico.

- `PROVINCIA`. Provincia a la que pertenece la población del centro. Es un campo alfanumérico.
- `COD_PROY`. Código del proyecto al que se hace la encuesta. Es un campo alfanumérico.
- `ID_EVALUACION`. Código identificativo de la evaluación, es un campo numérico y único.
- `FECHA DE EJECUCION`. Se trata de la fecha en la que se realizó la encuesta. Es un campo de tipo fecha.
- `COD_AUDITOR`. Código del auditor que realizó la encuesta. Se trata de un campo alfanumérico y único.
- `RESULTADO`. Se trata del resultado de la encuesta realizada. Es un campo numérico decimal entre 0 y 1.
- `TITULO_CUESTIONARIO`. Título que tiene asociado el cuestionario que se ha realizado. Se trata de un campo alfanumérico y único.

3.2 Análisis funcional y diagrama de arquitectura de flujo de datos

El sistema que se va a implementar consta de la siguiente funcionalidad:

- Primero recogerá los datos de la fuente y aplicará técnicas ETL, para así obtener un conjunto de datos válido y preparado para permitir a los usuarios finales la posiblidad de realizar el análisis de las encuestas realizadas en todos los centros de sus clientes y poder hacer un seguimiento de los mismos.
- Para poder realizar el análisis, el sisteman proporcionará distintos tipos de gráficos, como por ejemplo, la evolución temporal de los resultados obtenidos en cada uno de los cuestionarios; también la posiblidad de saber en qué centro se obtienen los mejores/peores resultados, entre otros.

Así pues, en la Figura 1 se muestra la arquitectura del flujo de datos.

Figura 1: Arquitectura del flujo de datos.

3.3 ¿Qué arquitectura de referencia usaría? Justifique la respuesta

Debido a que el sistema va dirigido a un departamento concreto de la compañía Mistery Shopping (al departamento antifraude) y haciendo uso de la Figura 2 en la que se muestran las diferencias entre el Data Warehose de Inmon y de Kimball, la arquitectura de referencia que se va a usar para implementar el sistema es la basada en el Data Warehouse de Kimball. Se toma esta decisión por las características propias del enunciado: se trata de una solución para un departamento concreto; el coste inicial es muy bajo y, debido a que este es un caso práctico de una asignatura, el tiempo de desarrollo no es muy elevado; y el equipo de desarrollo no tiene una especificación alta.

3.4 ¿Qué tecnología OLAP usaría? Justifique la respuesta

Dadas las características de la fuente de datos, expuestas en la sección 3.1, la tecnología OLAP que se va a usar va a ser ROLAP. Además de las características propias de la fuente de datos, se ha elegido ROLAP porque los usuarios no saben el tipo de análisis que van a realizar, por lo que al usar ROLAP se gana flexibilidad a la hora de la generación de análisis, además de que también permite poder realizar cualquier consulta ad-hoc sobre cualquiera de los atributos que contiene los datos.

	Inmon	Kimball
Presupuesto	Coste incial alto	Coste inicial bajo
Plazos	Requiere más tiempo de desarrollo	Tiempo de desarrollo inferior
Expertise	Equipo con especialización alta	Equipo con especialización media
Alcance	Toda la compañía	Departamentos individuales
Mantenimiento	Fácil mantenimiento	Mantenimiento más complejo

Figura 2: Diferencias entre el Data Warehouse de Inmon y de Kimball.

3.5 Si se utiliza ROLAP, ¿cuál de estos dos modelos se ajustaría mejor: modelo de estrella o el de copo de nieve?

La sencillez de la estructura de la fuente de datos supone que el modelo que mejor se ajustaría es el modelo de estrella, ya que utilizar un modelo de copo de nieve supondría añadir complejidad al sistema. En la Figura 3 de la sección 3.7 se muestra el modelo de estrella que se va a emplear para esta solución.

3.6 Si se utiliza ROLAP, hay que identificar y justificar si existe algún proceso de desnormalización de información que se deba realizar

Analizando los datos almacenados en la fuente de datos, se puede observar que no hay una normalización en los datos. Además, dado que se trata de encuestas que realiza la empresa, el sistema que proporciona la fuente de datos no es un sistema OLTP en el que se necesite rendimiento procesando pequeñas transacciones, por lo que la fuente de datos no está normalizada y, por consecuencia, no es necesario aplicar un proceso de desnormalización de la información.

3.7 Si se utiliza ROLAP, se debe incluir un diseño conceptual a modo explicativo junto con un diagrama

El diseño conceptual se basa en el modelo en estrella elegido en la sección 3.5 y se muestra en la Figura 3.

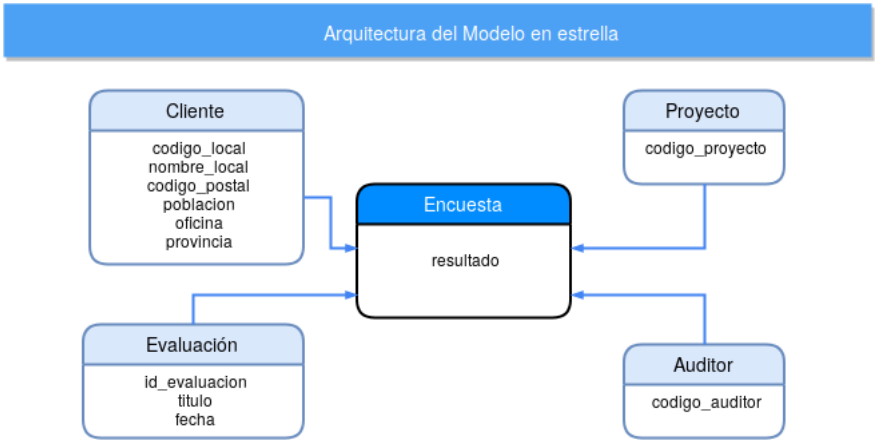


Figura 3: Modelo en estrella para el Data Mart Mistery Shopping.

3.8 Si se utiliza ROLAP, se debe incluir un diseño modelo lógico

Lo primero que se ha de hacer para realizar el diseño de un modelo lógico es identificar las dimensiones. Las **dimensiones** para este problema son: `Cliente`, `Proyecto`, `Auditor` y `Evaluación`.

Seguidamente, identificar los hechos. Los **hechos** son los resultados obtenidos en las encuestas.

Y por último, definir las métricas. Las **métricas** directas que se pueden definir son:

- *Resultado final de la encuestas, cuya función de agregación sería la suma.*
- *Encuesta con menor resultado, cuya función de agregación sería el mínimo.*
- *Encuesta con mayor resultado, cuya función de agregación sería el máximo.*
- *Número total de encuestas, cuya función de agregación sería el conteo.*

Una vez definidas las dimensiones, hechos y métricas, el siguiente paso es jerarquizar la información. Así pues, la jerarquización para todas las dimensiones anteriores es:

- `Cliente`. Esta dimensión tiene dos jerarquías:
  - `Local`, cuyos niveles son: `Código Local` y `Nombre Local`.
  - `Localización`, cuyos niveles son: `Provincia` y `Oficina`. A su vez, el nivel de `Provincia`, tiene los siguientes niveles:
    - `Población`.
    - `Código Postal`.
- `Proyecto`. Esta dimensión solamente tiene la jerarquía `Proyecto` cuyo nivel es `Código Proyecto`.
- `Auditor`. Esta dimensión solamente tiene la jerarquía `Auditor` cuyo nivel es `Código Auditor`.
- `Evaluación`. Esta dimensión solamente tiene la jerarquía `Evaluación` con los siguientes niveles:
  - `Evaluación`.
  - `Título`.
  - `Fecha`.

Así pues, en la Figura 4 se muestra el modelo lógico para el Data Mart Mistery Shopping.

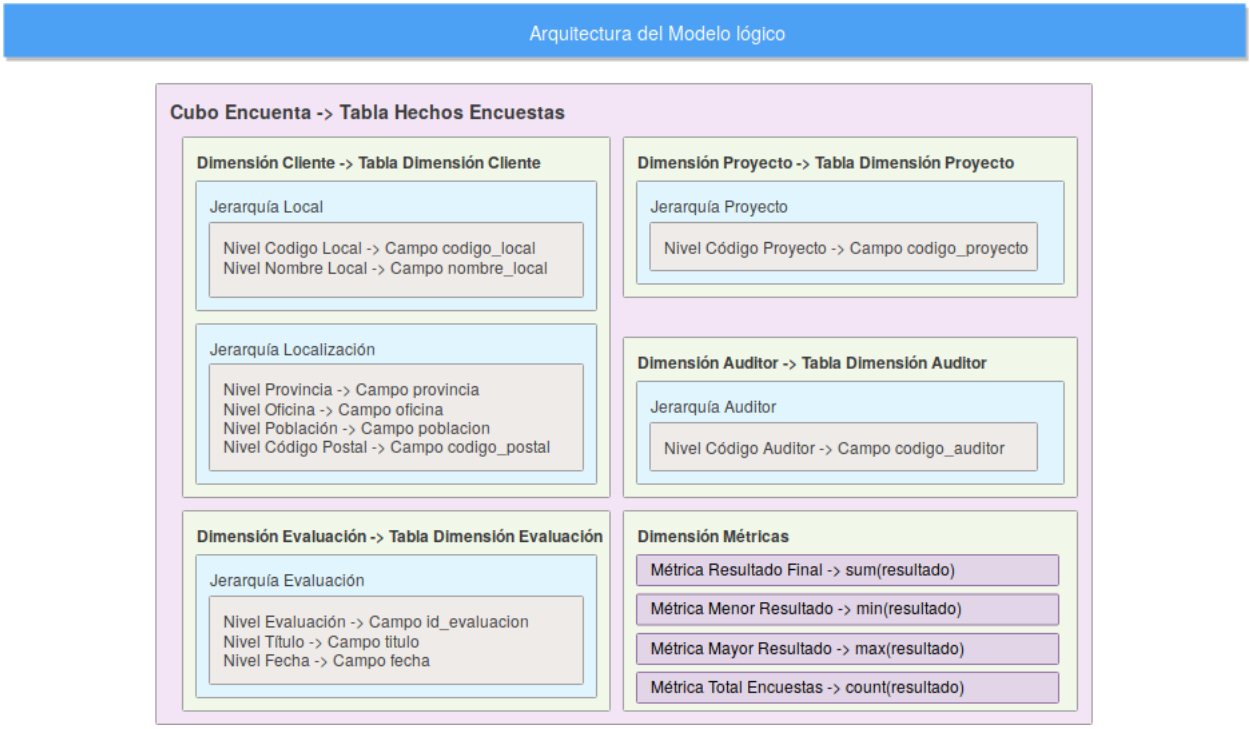


Figura 4: Modelo lógico para el Data Mart Mistery Shopping.

Partiendo del modelo en estrella de la Figura 3, la Figura 5 muestra el modelo lógico con los identificadores y claves de cada una de las tablas.

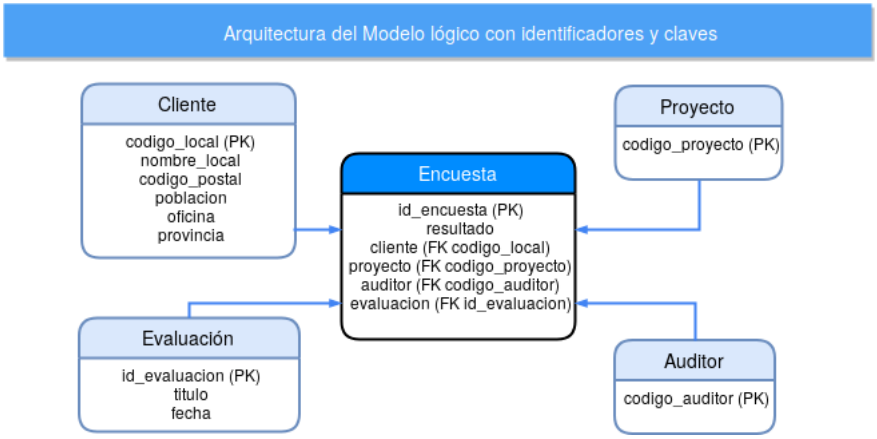


Figura 5: Modelo lógico con identificadores y claves para el Data Mart Mistery Shopping.

3.9 Si se utiliza ROLAP, se debe incluir un diseño modelo físico

En la Figura 6 se muestra el modelo físico que se va a emplear para este sistema:

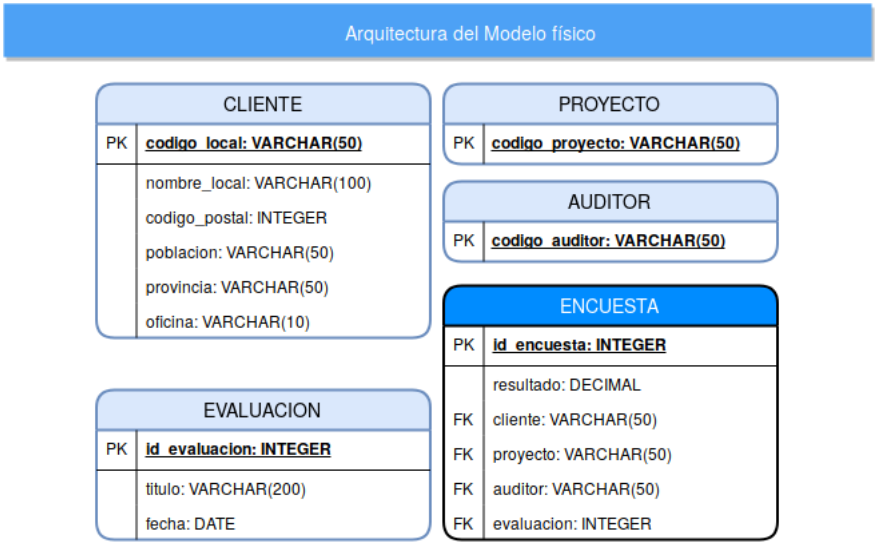


Figura 6: Modelo físico para el Data Mart Mistery Shopping.

3.10 Realizar la implementación del proceso ETL para generar y poblar el modelo multidimensional diseñado en los apartados anteriores

Para la realización de este apartado, se ha hecho uso de la herramienta Pentaho y se parte del JOB/Trabajo global `Global_IMF_def.kjb` proporcionado en el módulo de este caso práctico y que se encuentra situado en la carpeta `src/` y cuya estructura se muestra en la Figura 7. En la carpeta `src/` se encuentran todos los ficheros fuente utilizados para el proceso ETL.

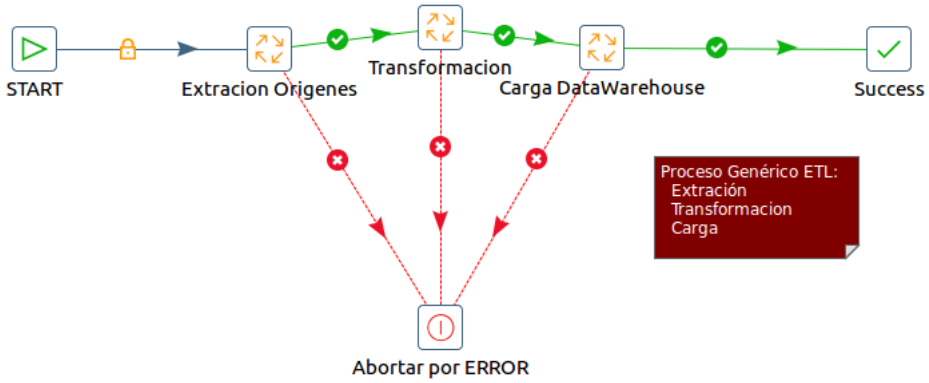


Figura 7: Estructura del trabajo `Global_IMF_def`.

A continuación se van a explicar en detalle cada uno de los trabajos en los que se divide este trabajo global:

3.10.1 Extracción Orígenes

La estructura general de este trabajo se encuentra en el fichero `Global_Extraccion.kjb` y se muestra en la Figura 8.

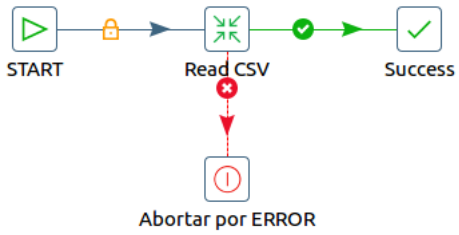


Figura 8: Estructura del trabajo `Global_Extraccion`.

En este trabajo lo único que se hace es leer los datos que vienen proporcionados por la fuente de datos descrita en la sección 3.1 y volcarlos todos en una tabla que servirá de entrada para el trabajo de Transformación explicado en la sección 3.10.2. Esto se hace a través de una transformación de Pentaho y que se encuentra en el fichero `Read_CSV.ktr` y cuya estructura se muestra en la Figura 9.

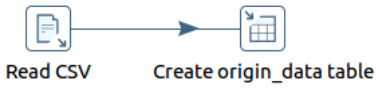


Figura 9: Estructura de la transformación `Read_CSV`.

3.10.2 Transformación

La estructura general de este trabajo se encuentra en el fichero `Global_Transformacion.kjb` y se muestra en la Figura 10.

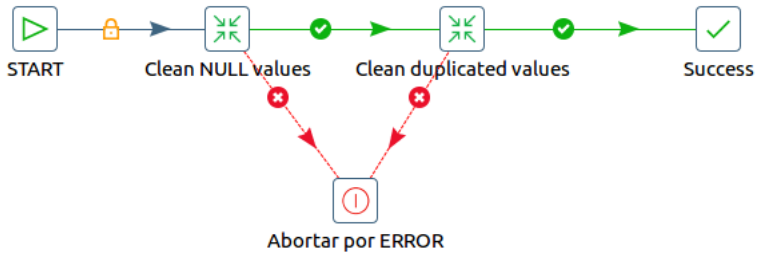


Figura 10: Estructura del trabajo `Global_Transformacion`.

En este proceso se realizan dos transformaciones: una para eliminar aquellos registros con valores nulos y otra para eliminar aquellos registros que contengan claves primarias duplicadas. A continuación se explica cada una con mayor detalle:

Eliminación valores nulos

Como se definió en el modelo físico en la sección 3.9, en todas las tablas existe un campo como clave primaria (`PK`) y por definición, este campo no puede ser nulo. Por lo que se van a eliminar aquellos registros que contengan en alguno de los campos `PK` un valor `NULL`. Para ello, se hace uso de la transformación `Clean_NULL_values.ktr` y cuya estructura se muestra en la Figura 11.

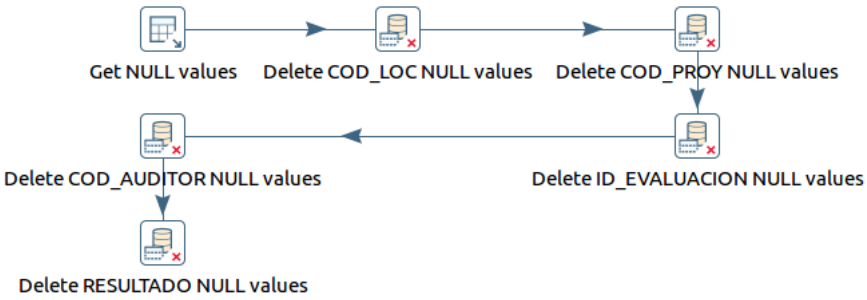


Figura 11: Estructura del trabajo `Clean_NULL_values`.

Además, para obtener estos registros, se hace uso de la siguiente sentencia SQL:

```sql
SELECT *
FROM origin_data
WHERE
        COD_LOC IS NULL OR
        COD_PROY IS NULL OR
        ID_EVALUACION IS NULL OR
        COD_AUDITOR IS NULL OR
        RESULTADO IS NULL
```

Importante destacar que aunque el campo `RESULTADO` es una clave primaria, se trata del campo que se quiere analizar por el departamento antifraude, por lo que este tampoco puede ser nulo.

Eliminación claves primarias duplicadas

Por definición, un valor de un campo que es clave primaria (`PK`) es único, ya que identifica de manera unívoca a cada registro. Pues siguiendo esta definición, se tiene que asegurar que en la fuente de datos no vienen registros distintos que hacen referencia a la misma clave primaria, lo cual duplicaría la clave primaria. Por ejemplo, supongamos que tenemos los siguientes registros, donde `codigo_local` es la clave primaria:

codigo_local	nombre_local
1	Pepito
1	Menganito

En este caso, si se intentan insertar ambos registros en la tabla `CLIENTE`, saltará un error diciendo que el valor `1` está duplicado. Como no podemos quedarnos con uno, porque no sabemos cuál es el correcto y cuál el erróneo, tenemos que eliminar ambos.

Siguiendo esta filosofía, se hace la transformación `Clean_duplicated_values.ktr` y cuya estructura se muestra en la Figura 12.

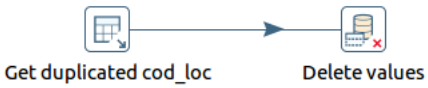


Figura 12: Estructura del trabajo `Clean_duplicated_values`.

Además, para obtener estos registros, se hace uso de la siguiente sentencia SQL:

```sql
SELECT *
FROM origin_data
WHERE COD_LOC IN (SELECT
                        COD_LOC
                  FROM origin_data
                  GROUP BY COD_LOC
                  HAVING
                        COUNT(DISTINCT NOMBRE_LOC) > 1 OR
                        COUNT(DISTINCT CP) > 1 OR
                        COUNT(DISTINCT PROVINCIA) > 1 OR
                        COUNT(DISTINCT POBLACION) > 1)
```

3.10.3 Carga DataWarehouse

Por último, tras realizar las transformaciones hay que realizar la carga del DataWarehouse. Para ello, se hace uso del trabajo `Global_Carga.kjb` y cuyo esquema se muestra en la Figura 13.

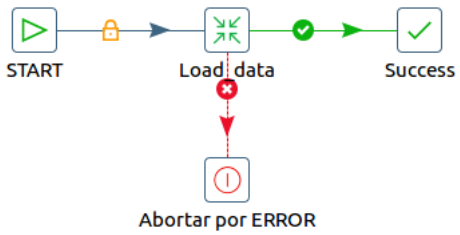


Figura 13: Estructura del trabajo `Global_Carga`.

Este trabajo lo que hace es crear cada una de las tablas definidas en la sección 3.9 con los datos que ya han sido transformados. Para ello, hace uso de la transformación `Load_data` cuyo estructura se muestra en la Figura 14

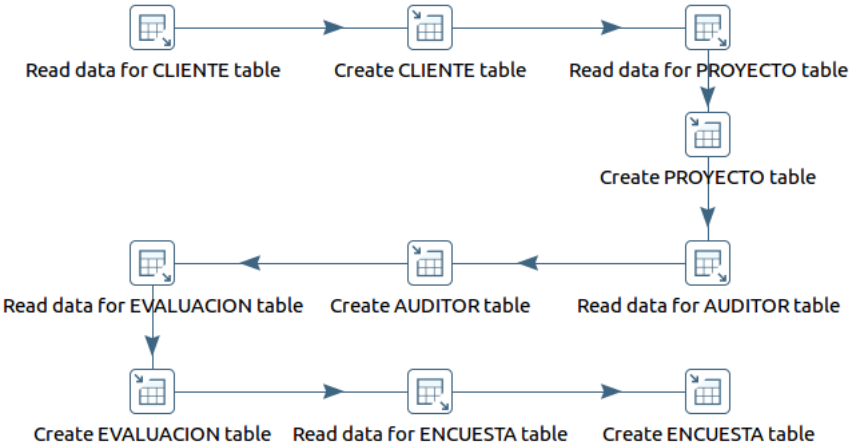


Figura 14: Estructura de la transformación `Load_data`.

Las sentencias SQL necesarias para crear cada una de las tablas se muestran a continuación:

```sql
CREATE TABLE CLIENTE
(
  CODIGO_LOCAL VARCHAR(50) PRIMARY KEY
, NOMBRE_LOCAL VARCHAR(200)
, CODIGO_POSTAL BIGINT
, POBLACION VARCHAR(50)
, PROVINCIA VARCHAR(50)
, OFICINA VARCHAR(10)
)
;


CREATE TABLE PROYECTO
(
  CODIGO_PROYECTO VARCHAR(50) PRIMARY KEY
)
;


CREATE TABLE AUDITOR
(
  CODIGO_AUDITOR VARCHAR(50) PRIMARY KEY
)
;

CREATE TABLE EVALUACION
(
  ID_EVALUACION BIGINT PRIMARY KEY
, TITULO VARCHAR(200)
, FECHA DATETIME
)
;


CREATE TABLE ENCUESTA
(
  ID_ENCUESTA BIGINT AUTO_INCREMENT NOT NULL PRIMARY KEY
, CLIENTE VARCHAR(50)
, PROYECTO VARCHAR(50)
, AUDITOR VARCHAR(50)
, EVALUACION BIGINT
, RESULTADO DOUBLE
)
;
```

Es importante destacar aquí que en la creación de la tabla `ENCUESTA` no se definen los campos `CLIENTE`, `PROYECTO`, `AUDITOR` y `EVALUACION` como claves foráneas (`FK`) porque al intentar lanzar el proceso entero y borrar todas las tablas con un `TRUNCATE` como hace Pentaho, da error por las restricciones de clave foránea, incluso si se pone `ON DELETE SET NULL` en la definición sigue dando el mismo error. Sin embargo, al rellenar la tabla, se hace uso de las claves primarias

definidas.

3.11 Implementación de modelo multidimensional diseñado en los puntos anteriores

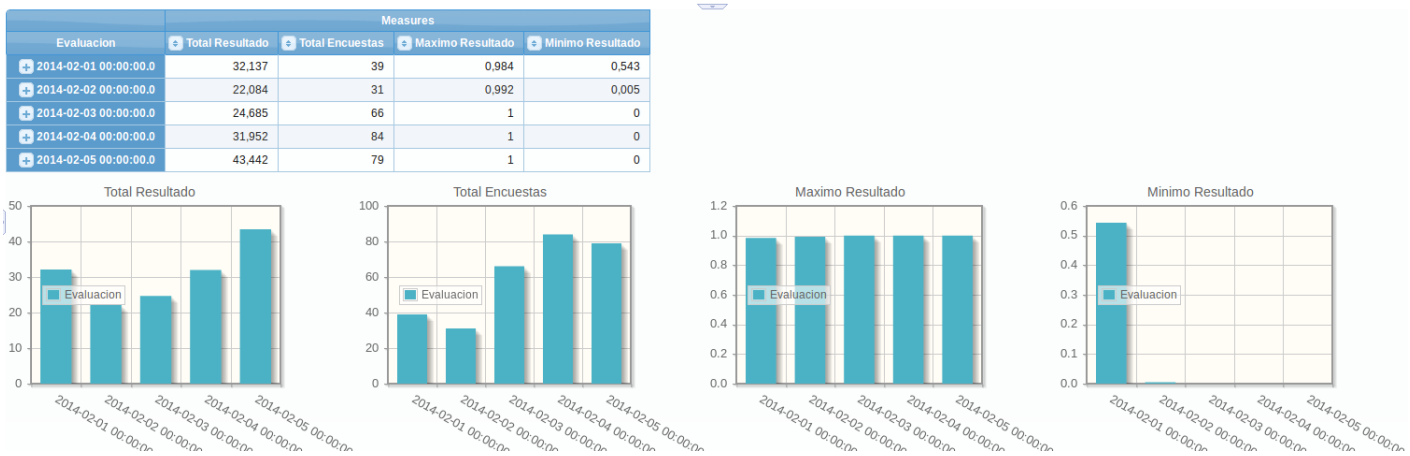
Para realizar la implementación del modelo multidimensional se ha hecho uso de la herramienta Wizard proporcionada y facilitada en el módulo. Con ella, y basándonos en el modelo lógico descrito en la sección 3.8, se ha implementado la implementación del modelo multidimensional que se muestra en la Figura 15.

Figura 15: Implementación del modelo multidimensional utilizando Wizard.

3.12 Análisis del modelo

En esta última sección se expone un análisis para probar el modelo multidimensional implementado en la sección 3.11, haciendo uso de los visores OLAP proporcionados en la máquina virtual del módulo.

Este análisis consiste en saber la cantidad de encuestas realizadas entre los días 1 (sábado) y 5 (miércoles) del mes de febrero del año 2014, así como el total de los resultados y el resultado máximo y mínimo obtenido por día. Gráficamente el resultado se muestra en la Figura 16, donde se puede observar que el número total de encuestas es menor durante el fin de semana (días 1 y 2), aunque sus resultados suelen ser mejores que los obtenidos en el resto de la semana (si nos fijamos, el resultado mínimo que se obtuvo el sábado día 1, fue entre 0.5 y 0.6, lo que nos puede indicar que los sábados se obtienen mejores resultados). También comentar que durante esos días, hubo al menos una encuesta que obtuvo un resultado cercano (incluso llegando) a 1 a la vez que también hubo encuestas que tuvieron como resultado 0.



Evaluacion	Measures			
	Total Resultado	Total Encuestas	Maximo Resultado	Minimo Resultado
2014-02-01 00:00:00.0	32,137	39	0,984	0,543
2014-02-02 00:00:00.0	22,084	31	0,992	0,005
2014-02-03 00:00:00.0	24,685	66	1	0
2014-02-04 00:00:00.0	31,952	84	1	0
2014-02-05 00:00:00.0	43,442	79	1	0

Figura 16: Análisis del modelo multidimensional utilizando visores OLAP.

Lógicamente, este análisis es muy básico, ya que era para probar el modelo multidimensional implementado. Las observaciones que se han obtenido y comentado anteriormente, no dejan de ser eso, observaciones, y que se deberían de confirmar analizando más profundamente el modelo y ver si lo que se ha comentado es real o no.

4 Conclusiones

Las conclusiones que se pueden obtener de este caso práctico es que se aleja, en mi opinión, bastante de lo enseñado en el módulo en ciertos puntos. Sin embargo, sí se acerca a lo que suele pedirse en el mundo real y eso es muy importante, ya que a los alumnos nos saca de nuestra zona de comfort estudiantil. Además, ayuda a comprender e interiorizar mejor los conceptos teóricos explicados en el módulo.

Personalmente, considero que este caso práctico ha sido uno de los que más me ha costado realizar, debido a que no estoy tan familiarizado este tema como puede ser la programación, de la que ha ido más de la mano el resto de módulos. He aprendido mucho realizando el caso práctico, y todo ha sido gracias al tutor, Jesús, que ha sabido ayudarme y guiarme desde el principio, cuando no sabía ni por dónde empezar.