IMF Business School

Universidad Camilo José Cela

Máster en Big Data & Business Analytics

Módulo V: Inteligencia de Negocio y Visualización

Documentación del caso práctico

22 de Octubre del 2018

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

Índice

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.

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.

	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.

3.4 ¿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.

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.

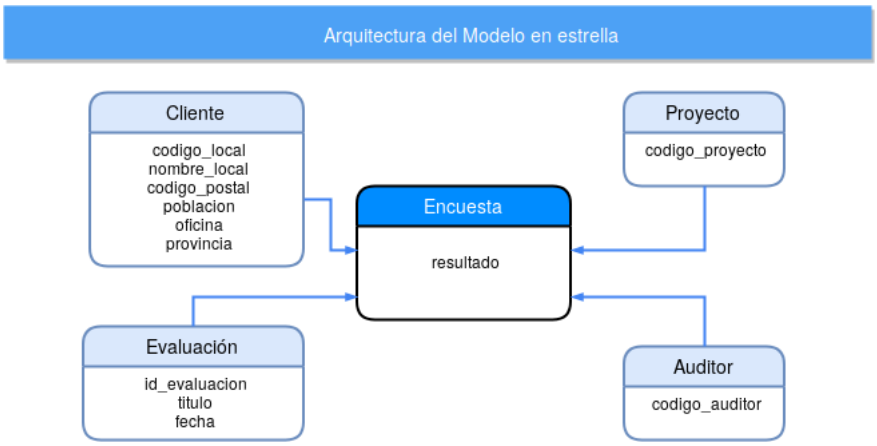

Figura 3: Modelo en estrella para el Data Mart Mistery Shopping.

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.*

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.

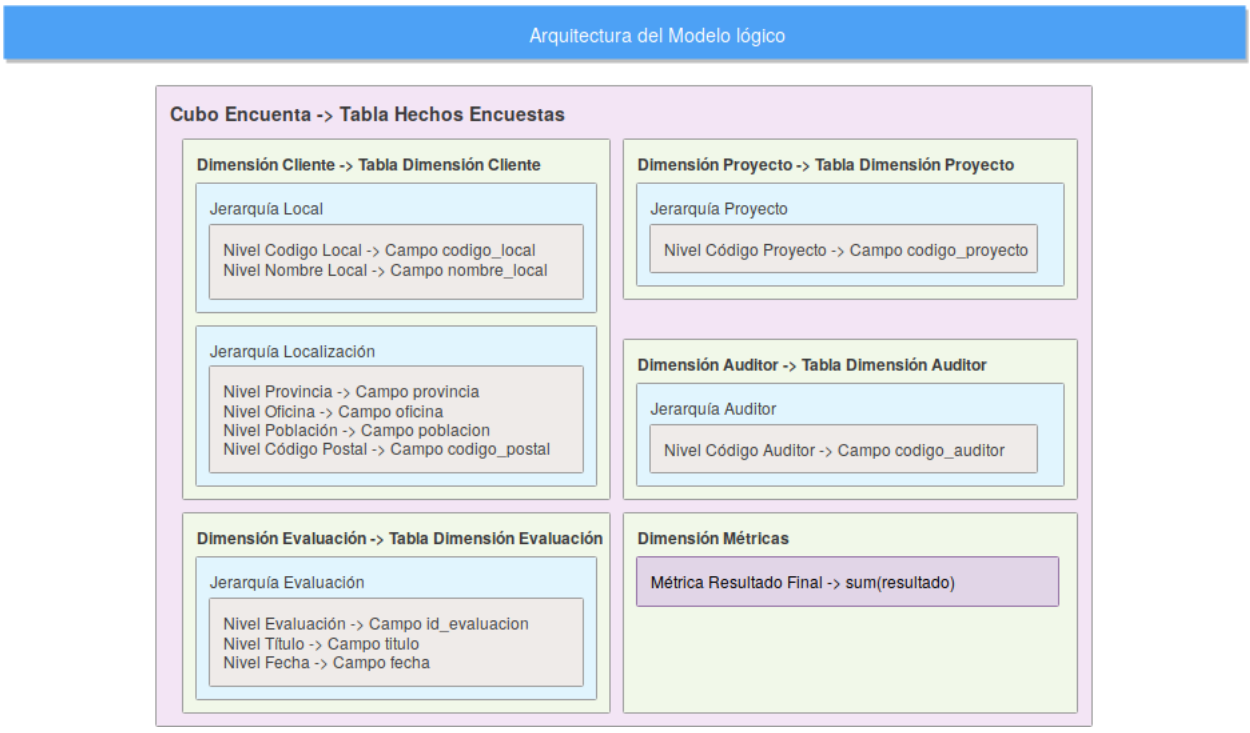

Figura 4: Modelo lógico para el Data Mart Mistery Shopping.

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.

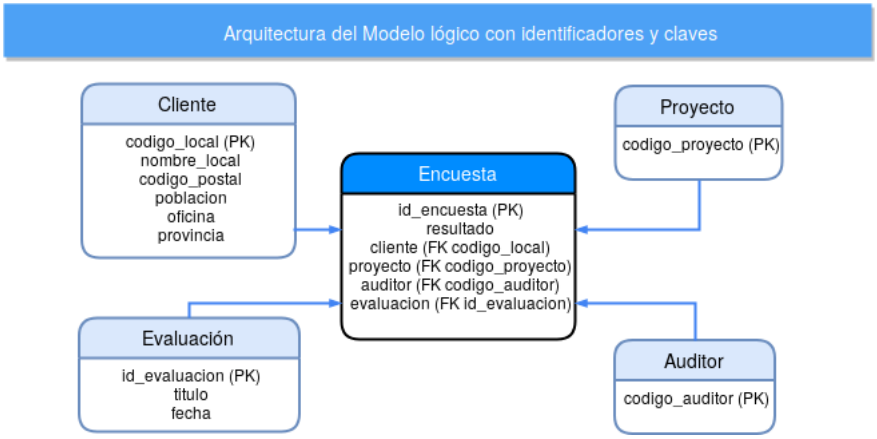

Figura 5: Modelo lógico con identificadores y claves para el Data Mart Mistery Shopping.

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:

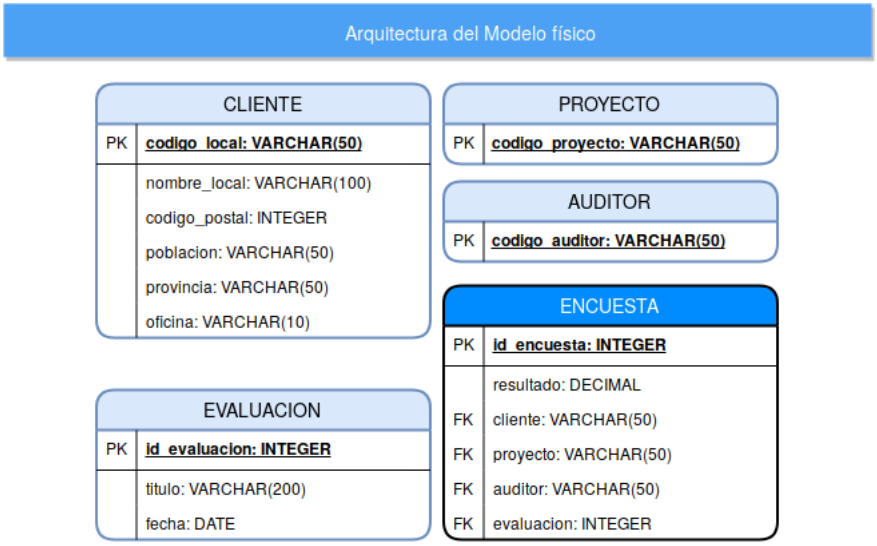

Figura 6: Modelo físico para el Data Mart Mistery Shopping.