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 del origen de datos (en este caso, del fichero CSV) y le aplicará técnicas ETL.
- Para posteriormente, formalizar un Data Mart con la información ya procesada.
- Se aplican herramientas OLAP para proporcionarle multidimensionalidad al Data Mart.
- Y así, 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.

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

Dada las características del análisis propuesto en la sección 3.2, la arquitectura que se va a emplear es la de datos en tres niveles:

- **Datos**. En este nivel se encontrarían el Origen de datos y el Data Warehouse de la Figura 1.
- **Aplicación**. En este nivel se encontrarían todas las funcionalidades de las herramientas OLAP.
- **Presentación**. Y en el último nivel se encontraría toda la parte de la Visualización de la Figura 1.

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.