CASO PRÁCTICO MÓDULO 5

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.

Se solicita :

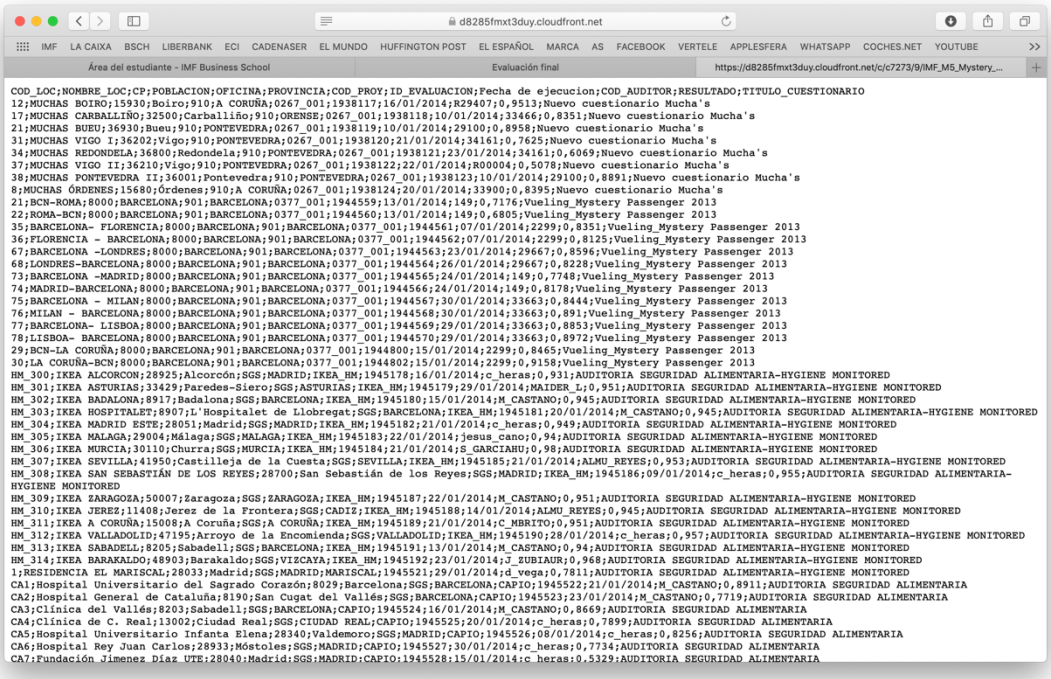
1. **Análisis de fuentes:**

1. Descripción global de las fuentes.

2. Descripción en detalle de cada campo.

3. Tipo de campo, naturaleza, cardinalidad aproximada.

La fuente de la que partimos es un archivo .csv con datos teóricos de visitas de un Mystery Shopper a establecimientos de diferentes empresas en los cuales los auditores han realizado un cuestionario que el departamento antifraude quiere valorar como cierto.



```
COD_LOC;NOMBRE_LOC;CP;POBLACION;OFICINA;PROVINCIA;COD_PROY;ID_EVALUACION;Fecha de ejecucion;COD_AUDITOR;RESULTADO;TITULO_CUESTIONARIO
12;MUCHAS BOIRO;15930;Boiro;910;A CORUÑA;0267_001;1938117;16/01/2014;R29407;0,9513;Nuevo cuestionario Mucha's
17;MUCHAS CARBALLIÑO;32500;Carballiño;910;ORENSE;0267_001;1938118;10/01/2014;33466;0,8351;Nuevo cuestionario Mucha's
21;MUCHAS BUEU;36930;Bueu;910;PONTEVEDRA;0267_001;1938119;10/01/2014;29100;0,8958;Nuevo cuestionario Mucha's
31;MUCHAS VIGO I;36202;Vigo;910;PONTEVEDRA;0267_001;1938120;21/01/2014;34161;0,7625;Nuevo cuestionario Mucha's
34;MUCHAS REDONDELA;36800;Redondela;910;PONTEVEDRA;0267_001;1938121;23/01/2014;34161;0,6069;Nuevo cuestionario Mucha's
37;MUCHAS VIGO II;36210;Vigo;910;PONTEVEDRA;0267_001;1938122;22/01/2014;R00004;0,5078;Nuevo cuestionario Mucha's
38;MUCHAS PONTEVEDRA II;36001;Pontevedra;910;PONTEVEDRA;0267_001;1938123;10/01/2014;29100;0,8891;Nuevo cuestionario Mucha's
8;MUCHAS ÓRDENES;15680;Órdenes;910;A CORUÑA;0267_001;1938124;20/01/2014;33900;0,8395;Nuevo cuestionario Mucha's
21;BCN-ROMA;8000;BARCELONA;901;BARCELONA;0377_001;1944559;13/01/2014;149;0,7176;Vueling_Mystery Passenger 2013
22;ROMA-BCN;8000;BARCELONA;901;BARCELONA;0377_001;1944560;13/01/2014;149;0,6805;Vueling_Mystery Passenger 2013
35;BARCELONA- FLORENCIA;8000;BARCELONA;901;BARCELONA;0377_001;1944561;07/01/2014;2299;0,8351;Vueling_Mystery Passenger 2013
36;FLORENCIA - BARCELONA;8000;BARCELONA;901;BARCELONA;0377_001;1944562;07/01/2014;2299;0,8125;Vueling_Mystery Passenger 2013
67;BARCELONA -LONDRES;8000;BARCELONA;901;BARCELONA;0377_001;1944563;23/01/2014;29667;0,8596;Vueling_Mystery Passenger 2013
68;LONDRES-BARCELONA;8000;BARCELONA;901;BARCELONA;0377_001;1944564;26/01/2014;29667;0,8228;Vueling_Mystery Passenger 2013
73;BARCELONA -MADRID;8000;BARCELONA;901;BARCELONA;0377_001;1944565;24/01/2014;149;0,7748;Vueling_Mystery Passenger 2013
74;MADRID-BARCELONA;8000;BARCELONA;901;BARCELONA;0377_001;1944566;24/01/2014;149;0,8178;Vueling_Mystery Passenger 2013
75;BARCELONA - MILAN;8000;BARCELONA;901;BARCELONA;0377_001;1944567;30/01/2014;3663;0,8444;Vueling_Mystery Passenger 2013
76;MILAN - BARCELONA;8000;BARCELONA;901;BARCELONA;0377_001;1944568;30/01/2014;33663;0,891;Vueling_Mystery Passenger 2013
77;BARCELONA- LISBOA;8000;BARCELONA;901;BARCELONA;0377_001;1944569;29/01/2014;33663;0,8853;Vueling_Mystery Passenger 2013
78;LISBOA- BARCELONA;8000;BARCELONA;901;BARCELONA;0377_001;1944570;29/01/2014;33663;0,8972;Vueling_Mystery Passenger 2013
29;BCN-LA CORUÑA;8000;BARCELONA;901;BARCELONA;0377_001;1944800;15/01/2014;2299;0,8465;Vueling_Mystery Passenger 2013
30;LA CORUÑA-BCN;8000;BARCELONA;901;BARCELONA;0377_001;1944802;15/01/2014;2299;0,9158;Vueling_Mystery Passenger 2013
HM_300;IKEA ALCORCON;28925;Alcorcón;SGS;MADRID;IKEA_HM;1945178;16/01/2014;c_heras;0,931;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_301;IKEA ASTURIAS;33429;Paredes-Siero;SGS;ASTURIAS;IKEA_HM;1945179;29/01/2014;MAIDER_L;0,951;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_302;IKEA BADALONA;8917;Badalona;SGS;BARCELONA;IKEA_HM;1945180;15/01/2014;M_CASTANO;0,945;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_303;IKEA HOSPITALET;8907;L'Hospitalet de Llobregat;SGS;BARCELONA;IKEA_HM;1945181;20/01/2014;M_CASTANO;0,945;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_304;IKEA MADRID ESTE;28051;Madrid;SGS;MADRID;IKEA_HM;1945182;21/01/2014;c_heras;0,949;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_305;IKEA MALAGA;29004;Málaga;SGS;MALAGA;IKEA_HM;1945183;22/01/2014;jesus_cano;0,94;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_306;IKEA MURCIA;30110;Churra;SGS;MURCIA;IKEA_HM;1945184;21/01/2014;S_GARCIAHU;0,98;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_307;IKEA SEVILLA;41950;Castilleja de la Cuesta;SGS;SEVILLA;IKEA_HM;1945185;21/01/2014;ALMU_REYES;0,953;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_308;IKEA SAN SEBASTIÁN DE LOS REYES;28700;San Sebastián de los Reyes;SGS;MADRID;IKEA_HM;1945186;09/01/2014;c_heras;0,955;AUDITORIA SEGURIDAD ALIMENTARIA-
HYGIENE MONITORED
HM_309;IKEA ZARAGOZA;50007;Zaragoza;SGS;ZARAGOZA;IKEA_HM;1945187;22/01/2014;M_CASTANO;0,951;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_310;IKEA JEREZ;11408;Jerez de la Frontera;SGS;CADIZ;IKEA_HM;1945188;14/01/2014;ALMU_REYES;0,945;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_311;IKEA A CORUÑA;15008;A Coruña;SGS;A CORUÑA;IKEA_HM;1945189;21/01/2014;C_MBRITO;0,951;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_312;IKEA VALLADOLID;47195;Arroyo de la Encomienda;SGS;VALLADOLID;IKEA_HM;1945190;28/01/2014;c_heras;0,957;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_313;IKEA SABADELL;8205;Sabadell;SGS;BARCELONA;IKEA_HM;1945191;13/01/2014;M_CASTANO;0,94;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
HM_314;IKEA BARAKALDO;48903;Barakaldo;SGS;VIZCAYA;IKEA_HM;1945192;23/01/2014;J_ZUBIAUR;0,968;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
1;RESIDENCIA EL MARISCAL;28033;Madrid;SGS;MADRID;MARISCAL;1945521;29/01/2014;d_vega;0,7811;AUDITORIA SEGURIDAD ALIMENTARIA-HYGIENE MONITORED
CA1;Hospital Universitario del Sagrado Corazón;8029;Barcelona;SGS;BARCELONA;CAPIO;1945522;21/01/2014;M_CASTANO;0,8911;AUDITORIA SEGURIDAD ALIMENTARIA
CA2;Hospital General de Cataluña;8190;San Cugat del Vallés;SGS;BARCELONA;CAPIO;1945523;23/01/2014;M_CASTANO;0,7719;AUDITORIA SEGURIDAD ALIMENTARIA
CA3;Clínica del Vallés;8203;Sabadell;SGS;BARCELONA;CAPIO;1945524;16/01/2014;M_CASTANO;0,8669;AUDITORIA SEGURIDAD ALIMENTARIA
CA4;Clínica de C. Real;13002;Ciudad Real;SGS;CIUDAD REAL;CAPIO;1945525;20/01/2014;c_heras;0,7899;AUDITORIA SEGURIDAD ALIMENTARIA
CA5;Hospital Universitario Infanta Elena;28340;Valdemoro;SGS;MADRID;CAPIO;1945526;08/01/2014;c_heras;0,8256;AUDITORIA SEGURIDAD ALIMENTARIA
CA6;Hospital Rey Juan Carlos;28933;Móstoles;SGS;MADRID;CAPIO;1945527;30/01/2014;c_heras;0,7734;AUDITORIA SEGURIDAD ALIMENTARIA
CA7;Fundación Jimenez Díaz UTE;28040;Madrid;SGS;MADRID;CAPIO;1945528;15/01/2014;c_heras;0,5329;AUDITORIA SEGURIDAD ALIMENTARIA
```

El archivo contiene 32797 registros con **12 campos diferentes**. A saber :

- **Código Local (COD_LOC) :**

  Campo de tipo string , con diferentes longitudes y/o caracteres ; en algunos casos solo contiene caracteres numéricos . Se propone limpieza de dicho campo.Tiene cardinalidad unívoca con los datos .
- **Nombre Local (NOMBRE_LOC)**

  Campo de tipo string , con diferentes longitudes y/o caracteres . Tiene cardinalidad unívoca con los datos .

- **Código Postal (CP)**

  Campo de tipo integer . Cardinalidad alta .

- **Población (POBLACION)**

  Campo de tipo string , con diferentes longitudes y/o caracteres . Tiene cardinalidad unívoca con los datos .

- **Oficina (OFICINA)**

  Campo de tipo string , con diferentes longitudes y/o caracteres . Tiene cardinalidad media .

- **Provincia (PROVINCIA)**

  Campo de tipo string , con diferentes longitudes y/o caracteres . Tiene cardinalidad baja .

- **Código Proyecto (COD_PROY)**

  Campo de tipo string , con diferentes longitudes y/o caracteres . Tiene cardinalidad alta .

- **Identidad Evaluación (ID_EVALUACION)**

  Campo de tipo integer. Tiene cardinalidad unívoca con los registros .

- **Fecha de Ejecución (Fecha de ejecución)**

  Campo de tipo Date . Tiene cardinalidad alta y formato dd/MM/yyyy.

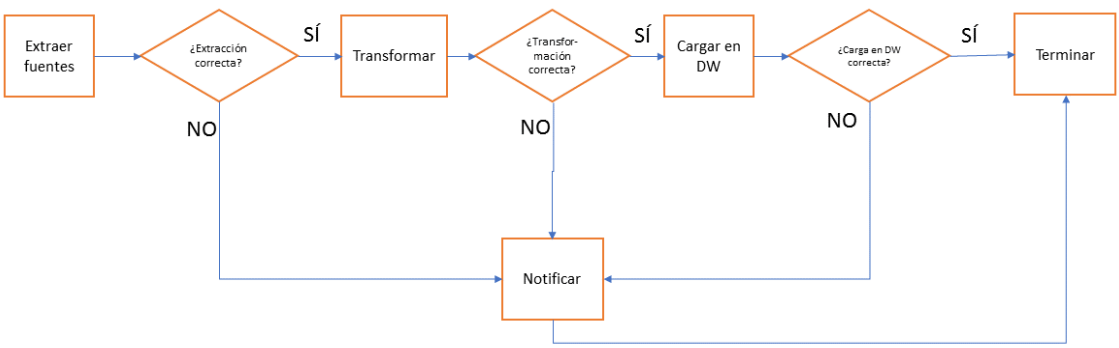
- **Código de Auditor (COD_AUDITOR)**

  Campo de tipo string , con diferentes longitudes y/o caracteres . Tiene cardinalidad alta .

- **Resultado (RESULTADO)**

  Campo de tipo numérico con longitud 6 y precisión estimada 4 . Tiene cardinalidad alta .

2. Análisis funcional y diagrama de arquitectura de flujo de datos.



3. ¿Qué arquitectura de referencia usaría? Justifique la respuesta.

Visto el origen de los datos , usaría una arquitectura multidimensional de Kimball , basada en Data Mart independientes puesto que el proceso es departamental y se centra en el análisis de un tipo de proceso de negocio (visitas a diferentes empresas por parte de un auditor).

4. ¿Qué tecnología OLAP usaría? Justifique la respuesta.

Aunque la tecnología R-OLAP o ROLAP (Relational On-line Analytical Processing) permite realizar análisis multidimensional dinámico a partir de los datos almacenados en una base de datos relacional que en nuestro caso no tenemos , en el modelo ROLAP los datos se almacenan como filas y columnas de forma relacional , que es el caso al que nos enfrentamos , parece lógico usar esta tecnología , mientras que la tecnología MOLAP guarda el dato en una estructura multidimensional directamente , algo que nosotros no tenemos .

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

Para implementar la capa física de nuestro problema , me parece más lógico el modelo más sencillo , esto es el modelo en estrella , cuyo esquema está definido por una sola tabla de hechos central, rodeada de tablas de dimensiones. En este caso, el esquema multidimensional no incluye ninguna jerarquía física, a nivel de base de datos relacional, entre los atributos de las diferentes dimensiones del esquema, de forma que cada una de ellas almacena todos sus atributos.

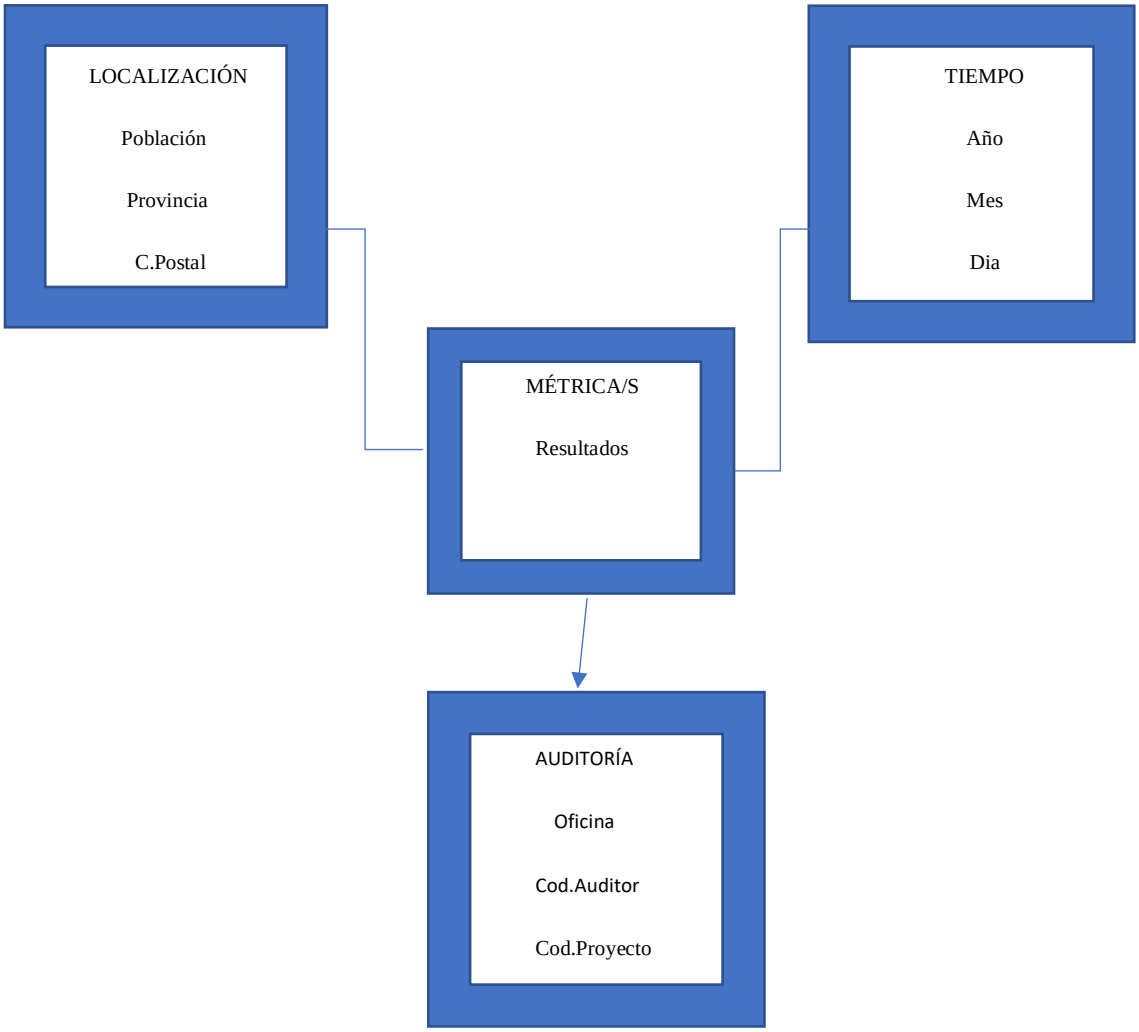
Descartaría , por tanto , el modelo de copo de nieve puesto que aunque al igual que el modelo de estrella, consta de una tabla central de hechos rodeada de tablas de dimensiones, cada nivel de la dimensión puede ser definido en otra tabla de dimensión y se conectan entre sí con una relación (n:1) , en nuestro caso , al ser cada hecho de la tabla un evento con diferentes niveles de dimensiones no relacionadas no parece adecuado .

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

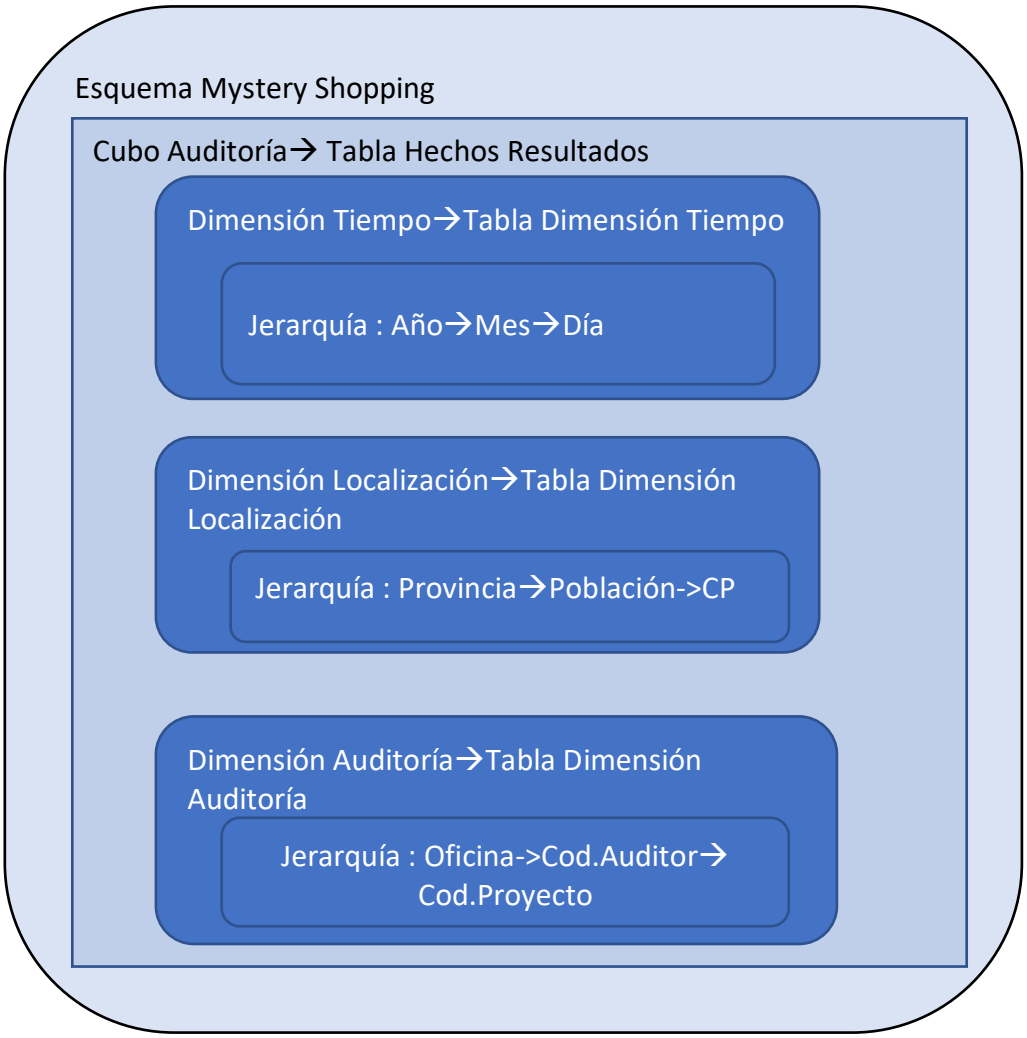
Como bien sabemos , en todo modelo de estrella los datos de las dimensiones deben ser almacenados de forma **desnormalizada** y, puesto que proceden de sistemas OLTP normalizados, es necesario aplicar un proceso de desnormalización, de forma que, como resultado final, se obtenga un modelo de estrella con una tabla de hechos y toda la información de dimensiones esté desnormalizada en tablas. En nuestro caso , puesto que los datos provienen de una única fuente no parece necesaria la desnormalización (más allá de la limpieza de datos de algunos campos de información como antes comentamos .

7,8 y 9. Si se utiliza ROLAP, se debe incluir un diseño conceptual a modo explicativo junto con diagramas físico y lógico.

Modelo Físico:

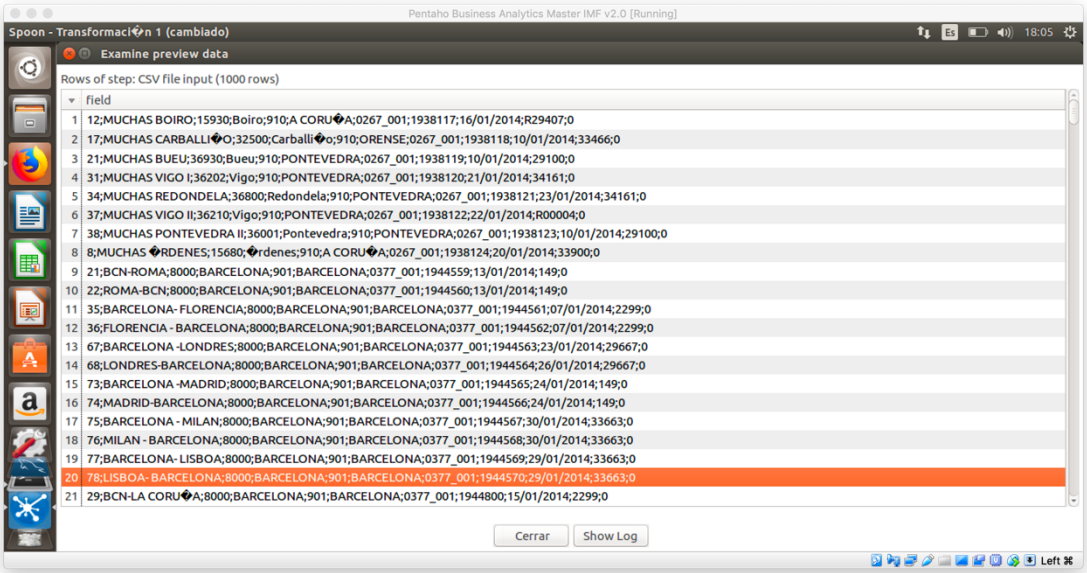


Modelo Lógico :

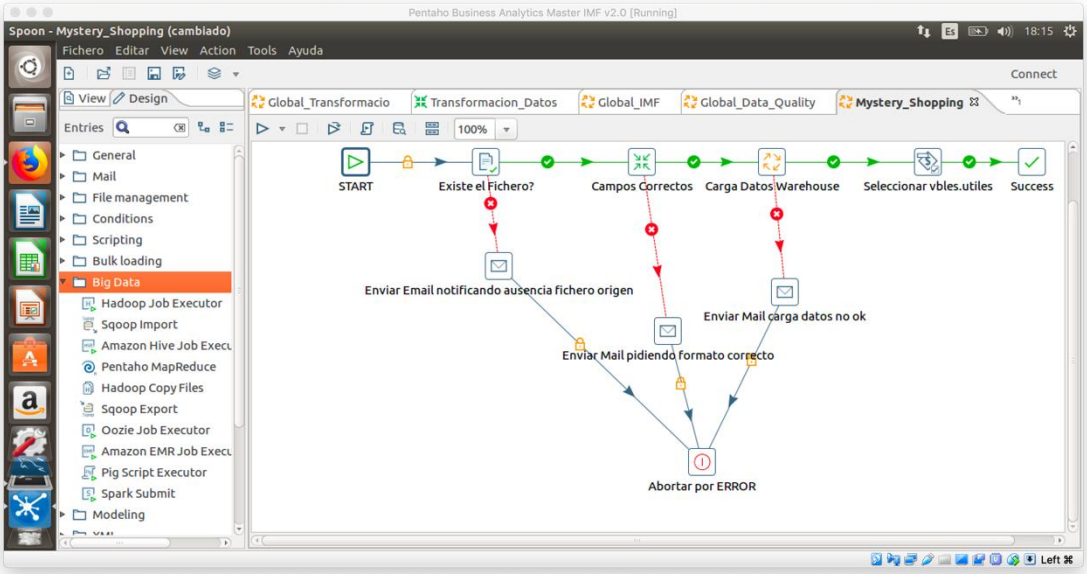


10. Realizar la implementación del proceso ETL para generar y poblar el modelo multidimensional diseñado en los apartados anteriores. Para ello, se partirá del JOB/Trabajo global "**Global_IMF.kjb**" que se puede descargar en el siguiente enlace: **Global_IMF.kjb.** Para la creación del DM/DW, hay que usar la base de datos MySql de la máquina virtual "**master_imf**".

Como primer paso previsualizamos el archivo de datos con Pentaho Business Analytics (se adjunta captura de pantalla) :

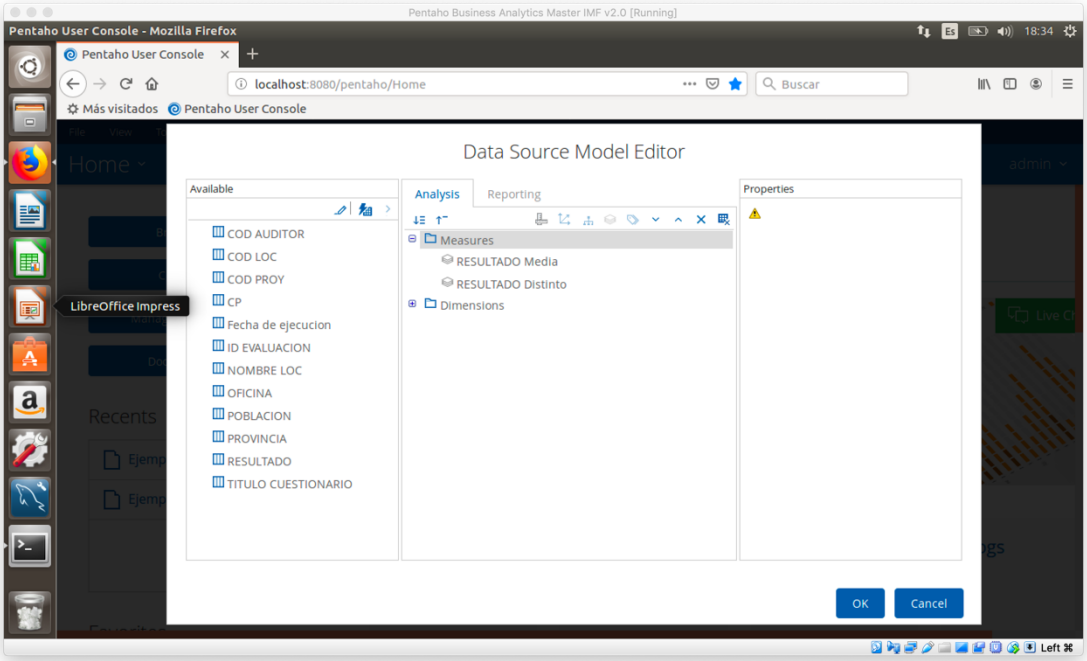


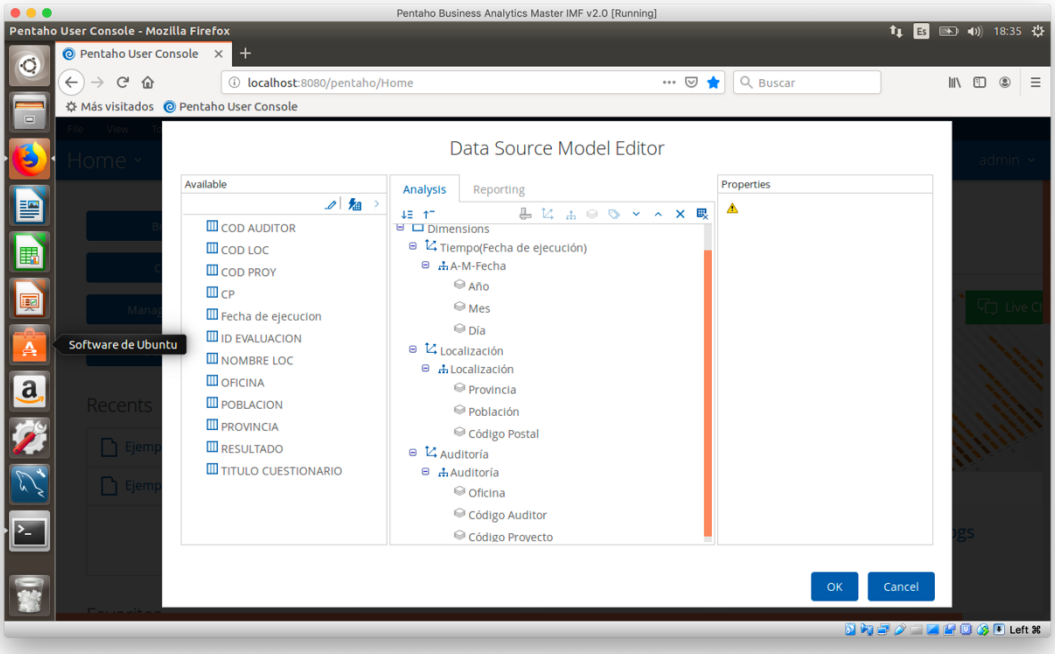
A partir de estos datos previsualizados , construímos el modelo ETL con la suite de Pentaho :



11.Implementación de modelo multidimensional diseñado mediante los puntos anteriores. Se debe realizar con la herramienta Wizard facilitada y mostrada en vídeos anteriores.

Adjuntamos dos capturas de pantalla del Wizard de Pentaho (PUC) en las que se pueden ver las métricas y las dimensiones del modelo (cubo multidimensional) creado con esta herramienta :





Se adjuntan también en el archivo zip los ficheros ***MysteryShopping.kjb*** y ***Modelo Wizard.png*** donde se pueden ver con más detalle los diagramas de arquitectura de flujo de datos así como como el Cubo de Datos con sus métricas , dimensiones , jerarquías y niveles generado mediante Pentaho Business Analytics Server .

12.Análisis de modelo. Se solicita realizar, al menos, un análisis, haciendo uso de un modelo multidimensional que refleje alguna situación relevante de ser explicada y comentada. Para ello, se hará uso de los visores OLAP disponibles en la MV.

En este punto he realizado los dos análisis que (con los datos que tenemos) me han parecido más relevantes a la hora de analizar el departamento antifraude si las visitas a los establecimientos realmente han sido llevadas a cabo . Estas son :

- Análisis de resultados de Oficina que lleva a cabo la visita frente a Provincia , el cual me parece relevante puesto que una oficina debería llevar a cabo visitas en un entorno cercano a su ubicación , por lo que algunas oficinas podrían resultar sospechosas (por ejemplo , la oficina 910 declara una visita a Córdoba muy alejada de su entorno en la zona Cantábrica …)

  Se adjunta el archivo pdf Oficina Vs Provincia con los resultados obtenidos mediante el visor OLAP Pilot 4j View .

- Análisis de resultados de Oficina que lleva a cabo la visita frente a Población , el cual me parece relevante puesto que una oficina debería llevar a cabo visitas en un entorno cercano a su ubicación y con criterios de auditoría similares , por lo que resultados demasiado buenos o demasiado malos deberían ser puestos en entredicho por el departamento antifraude .

  Se adjunta el archivo pdf Oficina Vs Población con los resultados obtenidos mediante el visor OLAP Pilot 4j View .