CASO PRÁCTICO FINAL

Maestría: Big Data y Ciencia de Datos

Módulo: Business Intelligence

Docente: Ing. Juan Carlos Benalcázar

Integrantes: Cabezas Gutierrez Cristhian Fabián

Calle Revelo Gustavo John

Gaibor Rojas Sonia Marlene

Muñoz Montenegro Miguel Fernando

Fecha: Octubre 2022

**Enunciado**

El objetivo de este caso práctico es poner en práctica los conceptos de inteligencia de negocio y *data warehousing* aprendidos a lo largo del módulo, haciendo hincapié en el diseño e implementación del modelo multidimensional y modelo de estrella, así como en los conceptos de dimensión, hechos, jerarquías, niveles, métricas y operaciones OLAP.

Del mismo modo, se pondrán en práctica los conocimientos adquiridos sobre inteligencia de negocio y ETL, prestando especial atención al diseño e implementación del proceso básico ETL hasta culminar en un modelo en estrella, así como en la implementación de dimensión y hechos en la capa física.

Los estudiantes deberán realizar:

* La implementación de un pequeño modelo multidimensional, haciendo uso de Pentaho Business Analytics Server o cualquier herramienta similar.
* La implementación de un pequeño proceso ETL, haciendo uso de Pentaho Data Integration (PDI) o cualquier herramienta similar.

Para la realización de este ejercicio práctico se partirá de los datos de mystery shopping **“IMF\_Mystery\_Shopping.csv”** disponible para descargar a continuación.

Todas las respuestas deben estar lo suficientemente desarrolladas y justificadas, independientemente de la contestación.

Se recomienda acompañar cada respuesta de todos los diagramas y representaciones que sea posible para argumentarla y justificarla.

ESCENARIO:

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, 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 o alguna herramienta similar, 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, Power BI o una herramienta similar, ha de realizarse una implementación ágil del modelo multidimensional.

El objetivo en este caso práctico 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 pide**

1. Análisis de fuentes:
   1. 1.Descripción global de las fuentes.
   2. 2.Descripción en detalle de cada campo.
   3. 3. Tipo de campo, naturaleza, cardinalidad aproximada.
2. Análisis funcional y diagrama de arquitectura de flujo de datos.
3. Generación de al menos 5 kpi's.
4. ¿Qué arquitectura de referencia se utilizaría? Justificar la respuesta.
5. ¿Qué tecnología OLAP se utilizaría? Justificar la respuesta.
6. Si se utiliza ROLAP, ¿Cuál de estos dos modelos se ajustaría mejor: el modelo en estrella o el de copo de nieve?
7. Si se utiliza ROLAP, identificar y justificar si existe algún proceso de desnormalización de información que se deba realizar.
8. Si se utiliza ROLAP, incluir un diseño conceptual a modo explicativo junto con un diagrama.
9. Si se utiliza ROLAP, incluir un diseño modelo lógico.
10. Si se utiliza ROLAP, incluir un diseño modelo físico
11. Realizar la implementación del proceso ETL para generar y poblar el modelo multidimensional diseñado en los apartados anteriores.
12. Implementación de modelo multidimensional diseñado mediante los puntos anteriores. Se puede realizar con la herramienta Power BI.
13. 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.
14. Generar un dashboard con los elementos más relevantes y que apuntalen a los kpi propuestos.

Para la creación del DW, se deberá usar cualquier base de datos RDBMS o NoSQL.

**FORMATO DE ENTREGA DEL CASO PRÁCTICO**

Se deberán entregar los siguientes archivos para la evaluación del caso práctico:

1. Memoria explicativa en formato .doc o .pdf, en la que se conteste a todas las preguntas solicitadas, incluyéndose todas las aclaraciones, justificaciones, comentarios y capturas de pantalla o imágenes que se consideren oportunas.
2. Archivos .kjb, .ktr, pbix, etc., necesarios para ejecutar la solución desarrollada.

La entrega de todos los archivos se realizará en un archivo comprimido en .rar o .zip.

1. **DESARROLLO**

El siguiente caso práctico se va a desarrollar en base al orden de los numerales planteados en el escenario:

1. **Análisis de fuentes de datos**

Se tiene como fuente de datos inicial al archivo IMF\_Mustery\_Shopping.csv que contiene ciertos datos de la Compañía Mystery Shopping ordenada por columnas. En términos generales este archivo contiene información de las evaluaciones realizadas en las diferentes provincias y localidades asignadas a un auditor con su respectivo resultado.

Haciendo un primer análisis de la fuente de datos referencial, se ha elaborado el siguiente diccionario de datos, que no es más que el conjunto de metadatos que enlistan las características lógicas y puntuales de los datos que, para el desarrollo de este caso práctico, se van a utilizar para las diferentes fases de la solución.

* 1. Descripción de detalle de cada campo

|  |  |
| --- | --- |
| Nombre del campo | Descripción |
| COD\_LOC | Guarda código de localidad |
| NOMBRE\_LOC | Guarda nombre de localidad |
| CP | Guarda código postal |
| POBLACION | Guarda nombre de población |
| OFICINA | Guarda código de oficina |
| PROVINCIA | Guarda nombre de provincia |
| COD\_PROY | Guarda código de proyecto |
| ID\_EVALUACION | Guarda id de evaluación |
| Fecha de ejecucion | Guarda fecha de ejecución |
| COD\_AUDITOR | Guarda código de auditor |
| RESULTADO | Guarda resultado de evaluación |
| TITULO\_CUESTIONARIO | Guarda nombre del cuestionario |

*Tabla 1. Detalle de los campos*

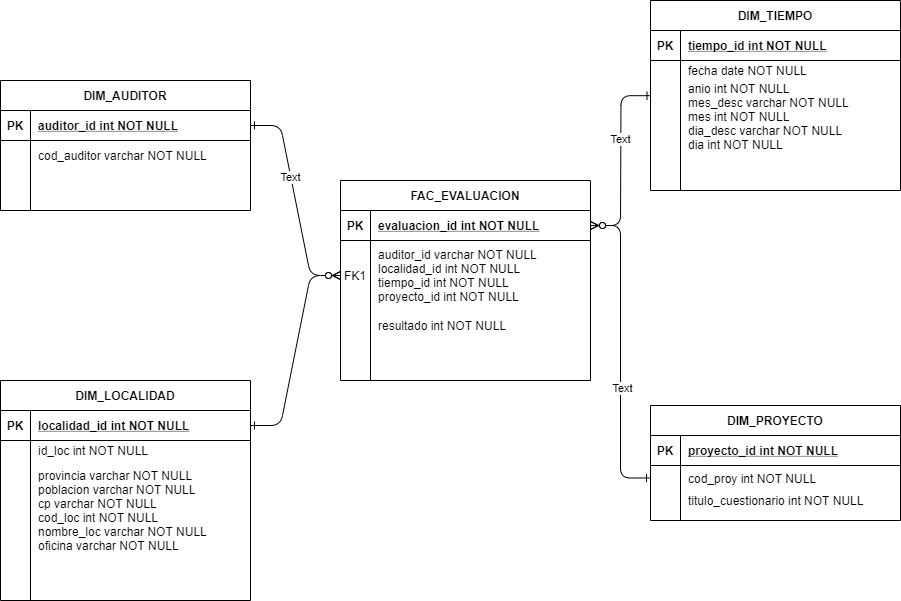
* 1. Tipo de campo, naturaleza

|  |  |
| --- | --- |
| Nombre del campo | Tipo de dato |
| COD\_LOC | varchar(15) |
| NOMBRE\_LOC | varchar(60) |
| CP | Number |
| POBLACION | varchar(50) |
| OFICINA | varchar(3) |
| PROVINCIA | varchar(25) |
| COD\_PROY | varchar(15) |
| ID\_EVALUACION | Number |
| Fecha de ejecucion | Date (DD/MM/YYYY) |
| COD\_AUDITOR | varchar(10) |
| RESULTADO | Number(5,4) |
| TITULO\_CUESTIONARIO | varchar(50) |

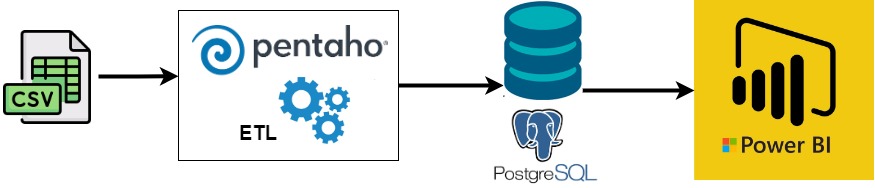
*Tabla 2. Detalle del tipo de campo*

* 1. Cardinalidad

**Cardinalidad**: 32797 registros.



1. **Análisis funcional y diagrama de arquitectura de flujo de datos**



En primera instancia, se realiza la extracción y carga de la fuente CSV en un staging, para posterior a ello, tomarla como fuente principal, para el ETL que permite generar las dimensiones, de localidad, proyecto, auditor y tiempo. Con esta primera etapa se pretende filtrar la información del staging unificándola con las dimensiones para finalmente cargar en la tabla de hechos evaluación. Todo esto en un ciclo de mejora continua de interpretación y evaluación del modelo. (**GUSTAVO**)

1. **Generación de al menos 5 KPI's.**

Tomando en consideración que el escenario plantea que el Departamento antifraude de la Compañía Mystey Shopping desea hacer un seguimiento y analizar la información relativa a las encuestas que se realizan en los distintos centros de clientes, se describen los siguientes KPI´s, siendo la definición de estos como “*Indicadores Clave de Desempeño que se componen de variables, factores o unidades de medida, que sirven para generar estrategias dentro de los departamentos de una empresa”*

a. Provincia donde el promedio del resultado de la encuesta es mayor a 0.97 por tipo de cuestionario de manera trimestral.

b. Población que tuvo los resultados más bajos o resultados más altos de la encuesta por fecha

c. Tipo de cuestionario aplicado por localidad en el año 2014.

d. Población que se le aplicó un determinado cuestionario en junio 2014.

e. Oficina aplicó cuestionarios y qué tipo en los años 2013 y 2014.

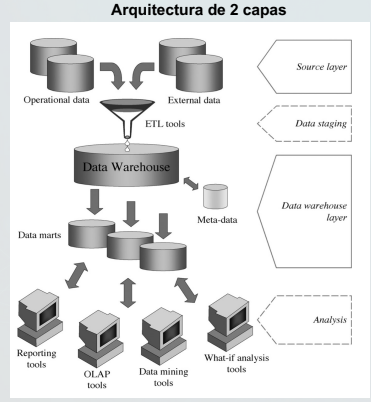
f. Número de encuestas por auditor y provincia.

g. Proyecto evaluado por localidad / provincia en los años 2013 y 2014.

(**MIGUEL**)

1. **Qué arquitectura de referencia se utilizaría?**

Arquitectura de referencia **(Gustavo)** quedó en graficar la arquitectura en base a esta) habría que describir en palabras también.



1. Qué tecnología OLAP se utilizaría?

Para este caso específico del diseño de la solución, describiremos a ambas tecnologías, OLAP utilizaría un servidor intermedio y en cambio ROLAP accede directamente a la base de datos relacional, por lo tanto, nuestra solución se basa en ROLAP que accederá a datos más directos, oportunos y actualizados que OLAP y también por tratarse de un escenario pequeño en dimensiones.

La diferencia básica entre OLAP y ROLAP es su arquitectura, es decir, las diferentes formas en que OLAP y ROLAP obtienen y acumulan información en una base de datos relacional.

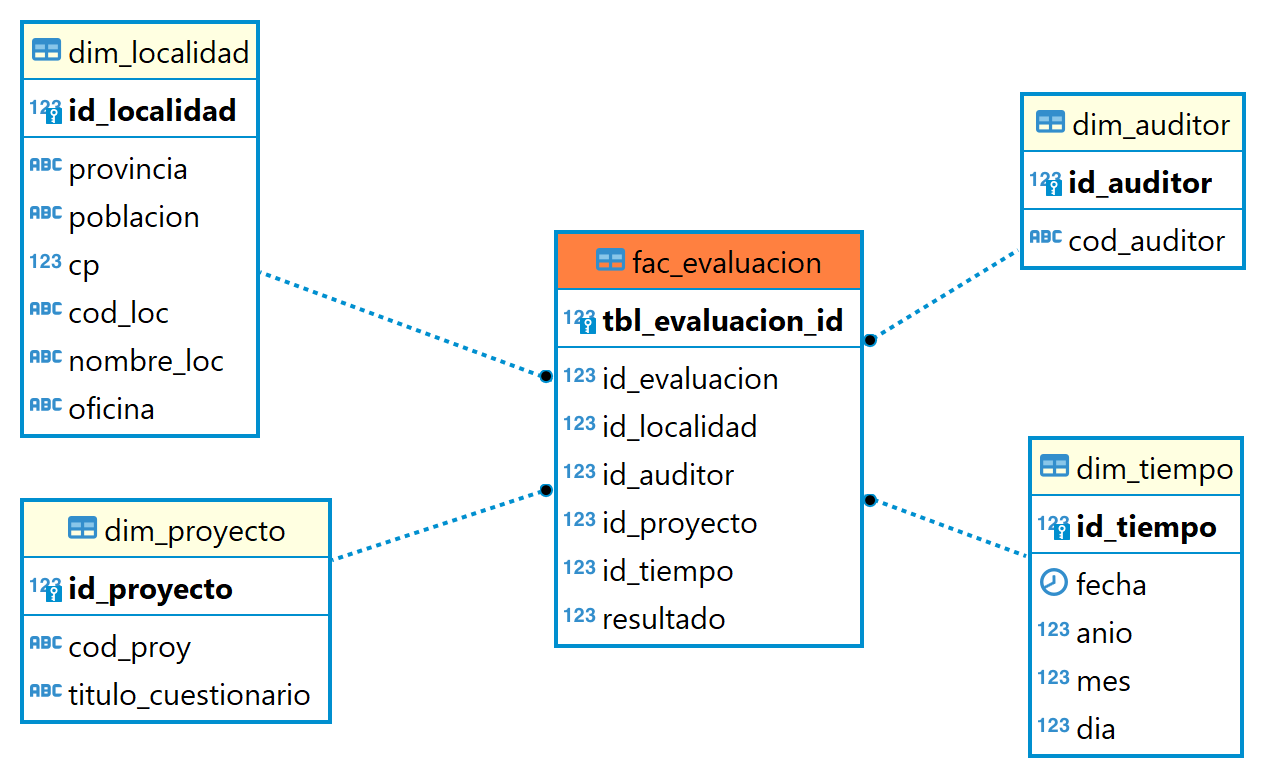
OLAP recoge los datos necesarios y los almacena en una memoria cache de datos especializada, esto permitiría un sólido rendimiento a un costo reducido. ROLAP ofrece una funcionalidad similar, pero realiza su análisis sobre la marcha, sin el paso intermedio de almacenar los datos en un servidor especial.

La ventaja radicaría en que permitirá a los usuarios realizar consultas rápidamente en bases de datos que son demasiado grandes para que un sistema OLAP pueda recorrerlas y analizarlas eficientemente.

1. **Si se utiliza ROLAP ¿Cuál de estos dos modelos se ajustaría mejor, estrella o copo de nieve?**

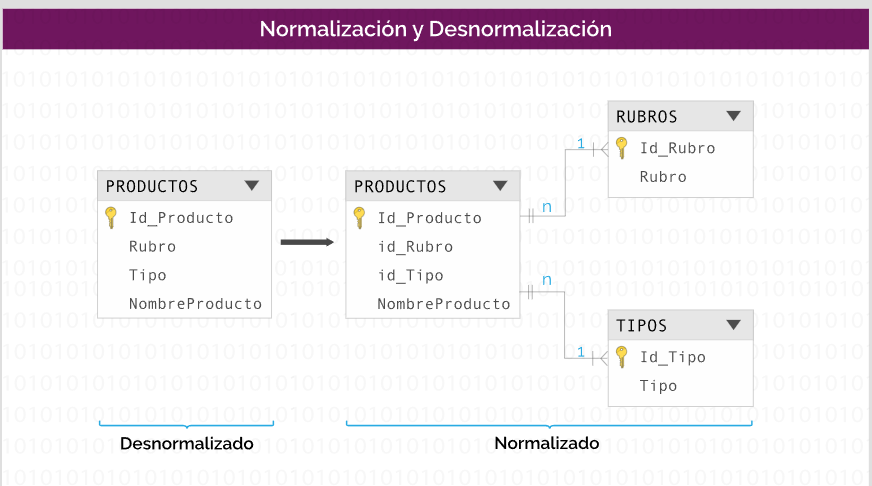
Para la solución se utilizará ROLAP (OLAP sobre relacional), ya que se va a usar una base de datos relacional para implementar la capa física y una capa de metadatos para implementar el modelo lógico multidimensional.

Se plantea un Modelo en **Estrella** para la solución, donde se tendrá una Tabla de Hechos denominada EVALUACION en el centro y sus puntas serán las dimensiones LOCALIDAD, AUDITOR, TIEMPO y PROYECTO.



1. **Si se utiliza ROLAP, identificar y justificar si se necesitaría un proceso de desnormalización que se deba realizar.**

La *desnormalización* es la duplicación intencionada de columnas en varias tablas y esto aumenta la redundancia de datos. (**GUSTAVO**)



1. **Si se utiliza ROLAP, incluir un diseño conceptual a modo explicativo junto con un diagrama**

**Fases:**

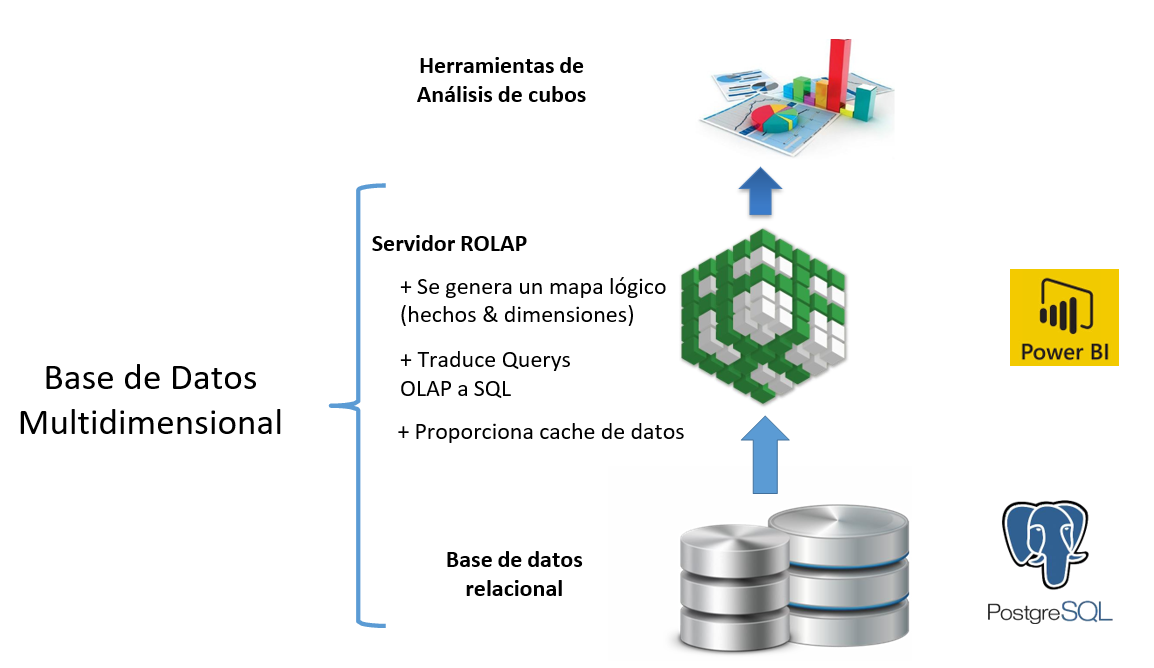
**1. Análisis de Requerimientos:** en esta sección se ha levantado los KPI´s de acuerdo a las necesidades del departamento antifraude de la compañía Mystery Shopping, con ello se tendría un modelo conceptual.

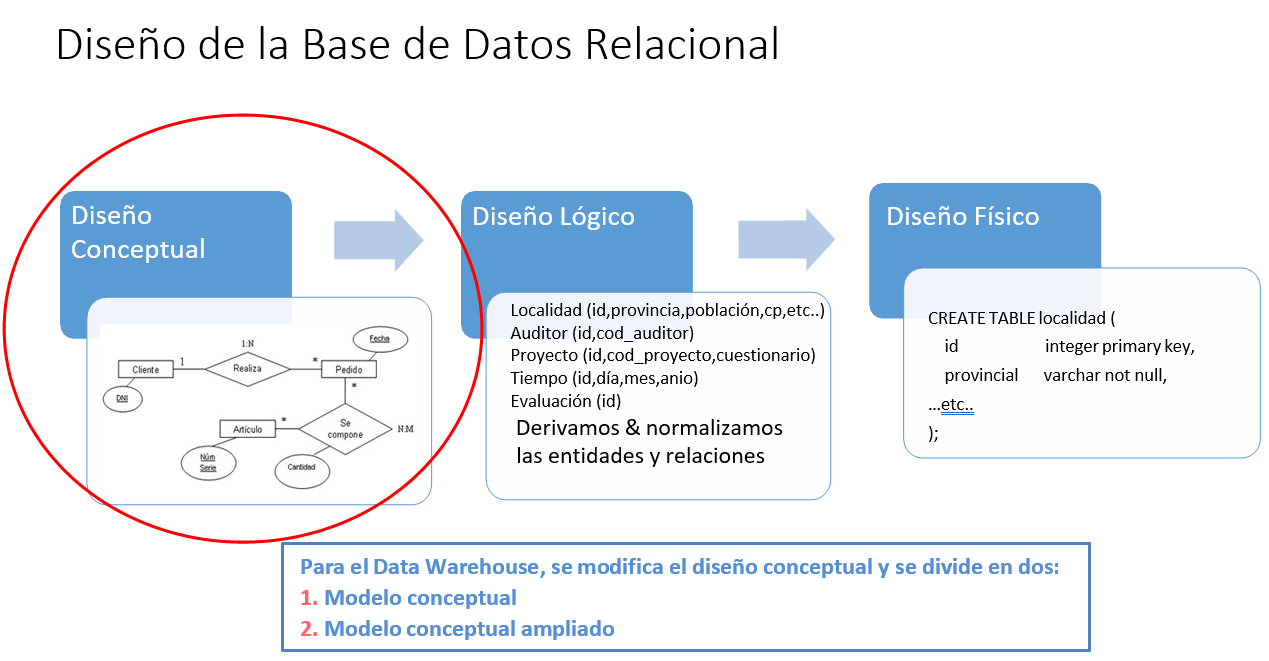
**2. Procesamiento Analítico OnLine Relacional (ROLAP)**: se va a analizar el procesamiento de las transacciones relacional para establecer la correspondencia entre el modelo conceptual creado en el paso anterior y las fuentes de datos. Aquí se va a determinar qué campo se incluirá en cada perspectiva y se va a ampliar el modelo conceptual con la información obtenida en esta fase.

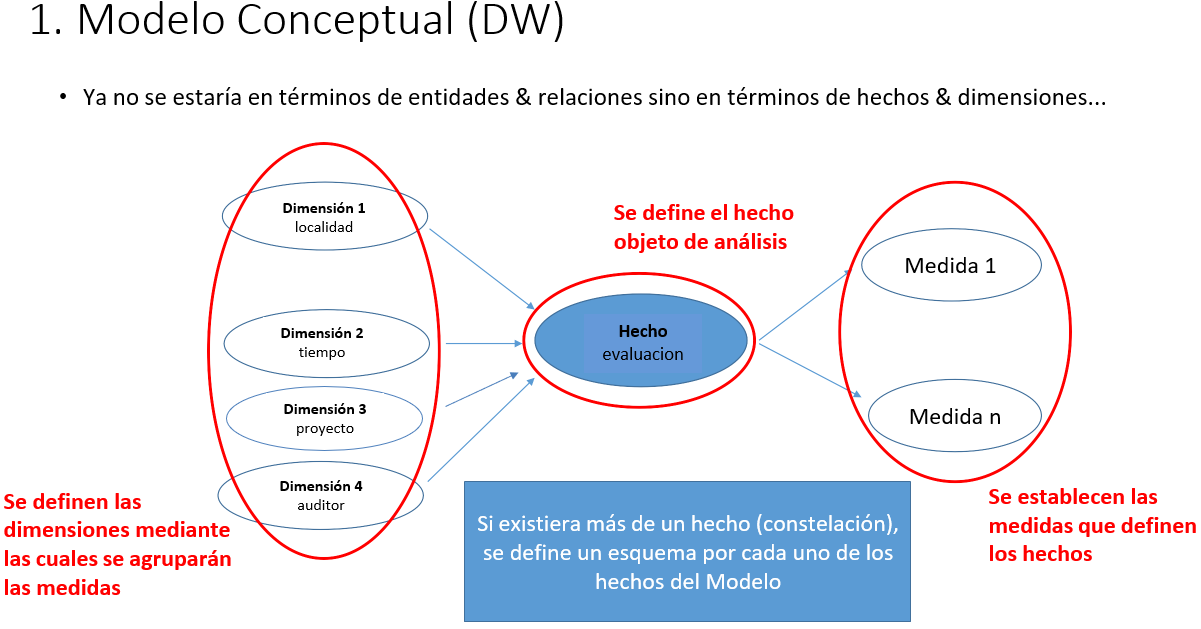
**3. Modelo Lógico del Data Warehouse:** en esta fase vamos a diseñar el Modelo Lógico mediante el análisis del tipo de modelo que se va a construir, en nuestro caso será el modelo estrella que tiene relación directa con la tabla de hechos y dimensiones pero carece de jerarquías.

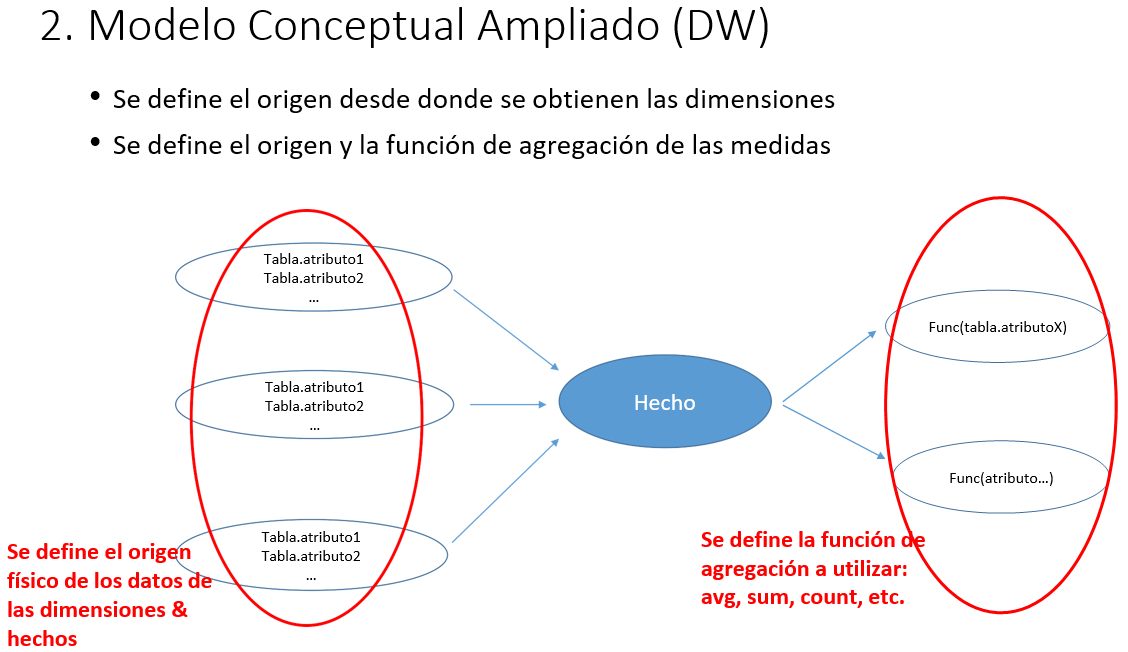
**4. Modelo Físico del Data Warehouse:** en esta última fase en donde se va a proceder a generar el modelo lógico del Data Warehouse.

Esquema de ROLAP a ser utilizado:









1. **Si se utiliza ROLAP, incluir un diseño del modelo lógico**

**Modelo Lógico del Data Warehouse:**

En esta fase se definió el modelo lógico del Data Warehouse, con las tablas de dimensiones y las de hecho, los objetos que se definieron son los nombres de las tablas con las claves primarias y foráneas en cada una de las tablas, así como las longitudes de los campos.

El modelo estrella se ha definido como el más adecuado para el análisis de los datos, en donde la tabla **evaluacion** tiene relación con todas las tablas.

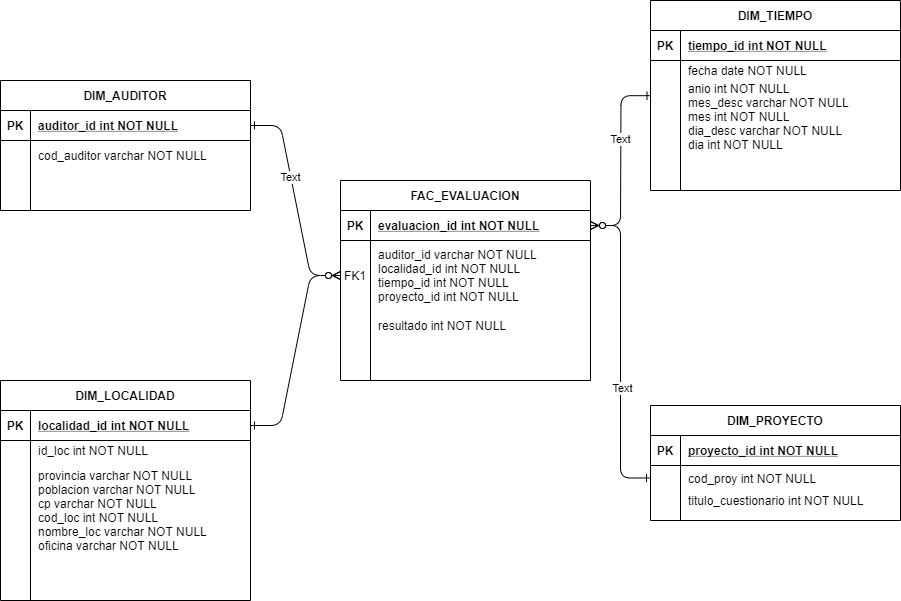
En el siguiente gráfico se aprecia el modelo lógico de la tabla evaluacion, en donde existe la relación de esta con las demás tablas de localidad, tiempo, proyecto y auditor.



1. **Si se utiliza ROLAP, incluir un diseño del modelo físico**

**Modelo Físico del Data Warehouse**

En el siguiente gráfico se aprecia el modelo físico el cual nos permite relacionar la tabla de hecho evaluación con las tablas de dimensiones.



1. **Realizar la implementación del proceso ETL parea generar y poblar el modelo multidimensional diseñado en los apartados anteriores**

Text, chat or text message

Description automatically generated

* **Staging**

Graphical user interface, application

Description automatically generated

* **Carga de dimensiones**

Diagram

Description automatically generated

* Dimensión Localidad

Chart, diagram, box and whisker chart

Description automatically generated

* Dimensión Proyecto

Chart, box and whisker chart

Description automatically generated

* Dimensión Auditor

A screenshot of a computer

Description automatically generated with low confidence

* Dimensión Tiempo

A picture containing text, sky, map

Description automatically generated

* **Carga de la tabla de hechos – Evaluación**

**Graphical user interface, application, Word

Description automatically generated**

1. **Implementación del modelo multidimensional diseñado mediante los puntos anteriores. Se puede realizar con una herramienta Power BI**

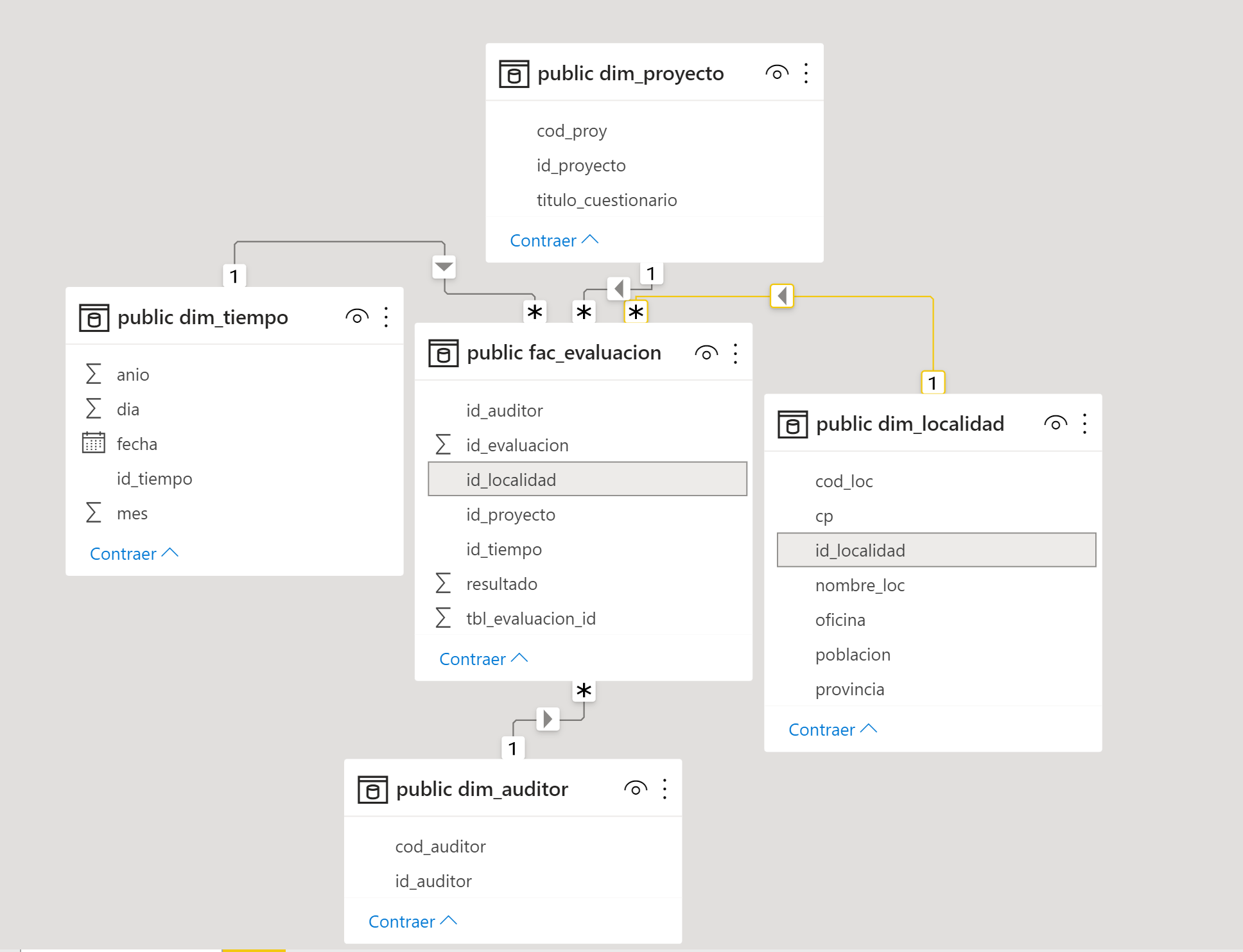
**Graphical user interface

Description automatically generated**

**Power BI Cómo se hizo la integración del ETL con el Power BI - (Cristian)**

**Capturas de pantalla del Power BI**

1. **Análisis del modelo. Se solicita, al menos un análisis haciendo uso de un modelo multidimensional que refleje alguna situación relevante de ser explicada y comentada.**

****

**Explicar el modelo escogido, los elementos y algo relevante que debamos mencionar del modelo multidimensional (Cristian)**

**Capturas de pantalla del Power BI**

1. **Generar un dashboard con los elementos más relevantes y que apunten a los kpi expuestos.**

Mediante la herramienta Power BI

1. Provincia donde el promedio del resultado de la encuesta es mayor a 0.97 por tipo de cuestionario de manera trimestral.

Chart, bar chart

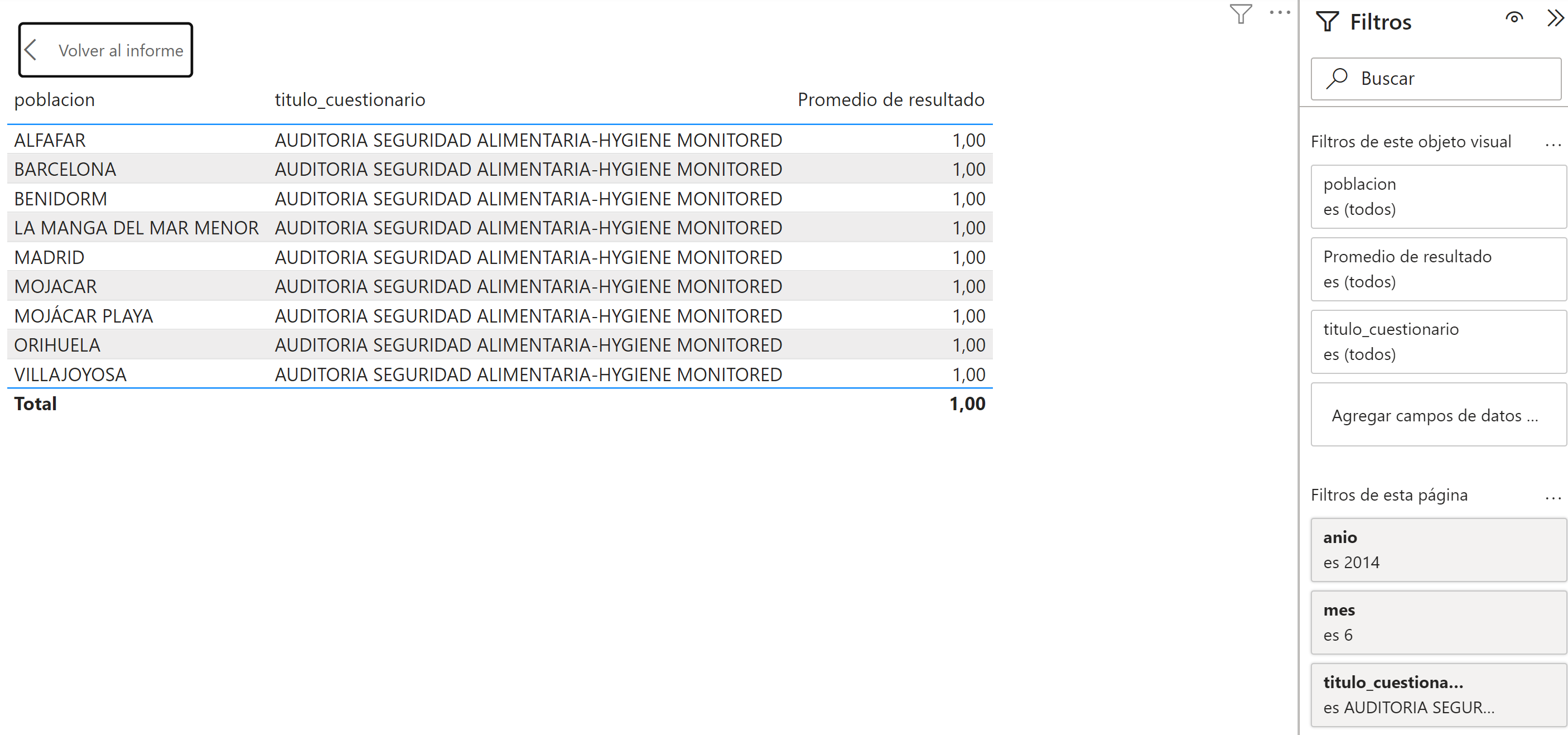
Description automatically generated

1. Población que tuvo los resultados más bajos o resultados más altos de la encuesta por fecha

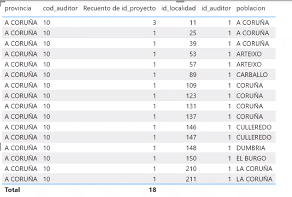
Graphical user interface, application

Description automatically generated

d. Población que se le aplicó un determinado cuestionario en el mes de junio y año 2014.



f. Número de encuestas por auditor y provincia.



g. Proyecto evaluado por localidad / provincia en los años 2013 y 2014

