

Caso práctico: Aplicación de los conceptos de Business Intelligence y Data Warehousing (Grupo L - Noviembre 2023)

Integrantes: Slin Castro, Dessiré Canchigre, Pablo Heredia, Verónica Alarcón

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 y su herramienta wizard para modelado multidimensional o puede utilizar cualquier otra herramienta para realizar el modelado.
- La implementación de un pequeño proceso ETL, haciendo uso de Pentaho Data Integration (PDI) o cualquier otra herramienta que cumpla las mismas funciones.

Para la realización de este ejercicio práctico se partirá de los datos de mystery shopping “IMF_Mystery_Shopping.csv” disponible para descargar.

Del mismo modo, se recomienda el uso de la máquina virtual del módulo para la ejecución del ejercicio.

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 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, 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 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:
 - a) Descripción global de las fuentes.
 - b) Descripción en detalle de cada campo.
 - c) Tipo de campo, naturaleza, cardinalidad aproximada.
2. Análisis funcional y diagrama de arquitectura de flujo de datos.
3. ¿Qué arquitectura de referencia se utilizaría? Justificar la respuesta.
4. ¿Qué tecnología OLAP se utilizaría? Justificar la respuesta.
5. Si se utiliza ROLAP, ¿cuál de estos dos modelos se ajustaría mejor: el modelo en estrella o el de copo de nieve?
6. Si se utiliza ROLAP, identificar y justificar si existe algún proceso de desnormalización de información que se deba realizar.
7. Si se utiliza ROLAP, incluir un diseño conceptual a modo explicativo junto con un diagrama.
8. Si se utiliza ROLAP, incluir un diseño modelo lógico.
9. Si se utiliza ROLAP, incluir un diseño modelo físico.

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 o trabajo global “Global_IMF.kjb” que se puede descargar desde la máquina virtual.
11. Implementación de modelo multidimensional diseñado mediante los puntos anteriores. Se debe realizar con la herramienta wizard facilitada y mostrada en vídeos anteriores.
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 máquina virtual.

Para la creación del DM/DW, se deberá usar cualquier base de datos relacional o NoSQL “master_imf”.

Formato de entrega del caso práctico: Archivos .kjb, .ktr, .pbix, .py, etc., necesarios para ejecutar la solución ETL desarrollada. en un archivo comprimido en .rar o .zip.

I. DESARROLLO

GENERALIDADES:

Todo el código de la solución también está disponible en este repositorio de [github](https://github.com/slincaastro/mysteryShopping) (<https://github.com/slincaastro/mysteryShopping>):

1. ANÁLISIS DE FUENTES:

1. Descripción global de las fuentes.

Para la ejecución de este ejercicio se cuenta con un fichero .csv con un tamaño de 260kb, contiene 32,797 filas y 12 columnas.

2. Descripción en detalle de cada campo.

Tabla 1: DESCRIPCIÓN DE CAMPOS

CAMPO	TIPO	DESCRIPCIÓN
COD_LOC	Texto, String, Alfanumérico, ASCII Extendido	Se infiere que es el identificador de una locación, no es un campo estandarizado, se encuentran patrones numéricos y alfanuméricos e incluso valores vacíos, la longitud va de 1-15 caracteres.
NOMBRE_LOC	Texto, String, Alfanumérico, ASCII Extendido	El campo contiene nombres de localizaciones, que varían de formato y estándar, se puede encontrar valores como : Estado-Ciudad, Código Postal - ciudad, Descripción - Ciudad y tiene una longitud de 2 a 52 caracteres

CP	Numérico, Integer	Es el código postal de la locación.
POBLACION	Texto, String, Alfanumérico, ASCII Extendido	Define la población en la que está el local, contiene algunos valores vacíos y los valores tienen caracteres entre 2 y 41
OFICINA	Numérico, Integer	Es un código numérico único de 3 dígitos que indica la oficina donde se hizo la evaluación
PROVINCIA	Texto, String, Alfanumérico, ASCII Extendido	Es un campo de texto, que tiene una longitud entre 4 y 22 caracteres e indica la provincia de origen de la tienda
COD_PROY	Texto, String	Es un campo que en su mayoría sigue un patrón definido #####_#### y los dígitos numéricos a los extremos de varían.
ID_EVALUACION	Numérico, Integer	Es un campo numérico único que podría ser definido como clave única ya que su cardinalidad es Alta.
FECHA DE EJECUCION	Date	Es la fecha de ejecución de la encuesta, que va desde 1995-03-05 hasta 2014-12-12.
COD_AUDITOR	texto, String, Alfanumérico, ASCII Extendido	Es un campo alfanumérico con diferentes patrones en diferentes casos, que indica el código del auditor que ejecutó la encuesta.
RESULTADO	Numérico, Punto Flotante	Es un número decimal entre 0 y uno que determina el riesgo de fraude que puede tener ese local.
TITULO_CUESTIONARIO	texto, String, Alfanumérico, ASCII Extendido	Es una cadena de caracteres, entre 5 y 50, que indica el título asignado a la encuesta.

Elaborado por: Grupo L

3. Tipo de campo, naturaleza, cardinalidad aproximada*.

Tabla 2: TIPOS DE CAMPOS

Nombre Campo	TIPO DE DATO	Cardinalidad (Datos Únicos)	Valor Máximo	Valor Mínimo	Longitud Máxima	Longitud Mínima
COD_LOC	String	5646	Z001651	0	15	1
NOMBRE_LOC	String	6701	♦vila El Boulevard	1001 - C/ Maestro Aguilar, nº 25	52	2
CP	Int	1797	99999	0	5	1
POBLACION	String	2029	♦viles-Asturias	--	41	2
OFICINA	Int	13	SGS	901	3	3
PROVINCIA	String	53	ZARAGOZA	A CORUÑA	22	4
COD_PROY	String	223	ZOUK	0001_077	14	2
ID_EVALUACION	Integer	32797	2016756	1938117	7	7
Fecha de ejecución	Date	188	nan	01/01/2014	10	3
COD_AUDITOR	String	1169	test	10	10	1

RESULTADO	float	4205	1	0	6	1
TITULO_CUESTIONARIO	String	439	prueba	ABBA SID MYST ERY	50	5

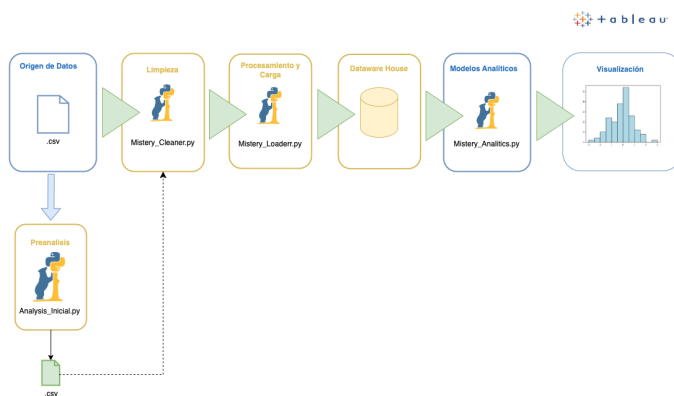
Elaborado por: Grupo L

*Datos obtenidos a partir de un script de análisis de datos en python adjunto, ([analisis_inicial.py](#))

2. ANÁLISIS FUNCIONAL Y DIAGRAMA DE ARQUITECTURA DE FLUJO:

Para el desarrollo de este punto, se propone la ejecución de los siguientes pasos: preanálisis, limpieza, procesamiento y carga, lectura, generación de modelos y, visualización; como lo muestra la ilustración 1 a continuación:

Ilustración 1: Diagrama de Arquitectura de Flujo



Elaborado por: Grupo L

Pre-Análisis: Esta etapa es importante ya que permite tener una primera vista de los datos a procesar, validar su calidad para la ejecución posterior de los procesos de Limpieza y estandarización, así como, para la ejecución de un proceso más limpio de análisis y aprovechamiento de los datos. Este primer acercamiento, permitió identificar los errores de la base de datos, entre los más recurrentes constan el código postal, celdas vacías y los formatos de fecha. Esta fase también permitió trazar las líneas de análisis de acuerdo con la información proporcionada y apuntar ideas de visualización de información teniendo en consideración líneas temporales, niveles de granularidad, entre otros.

Limpieza: Con los insights obtenidos del pre-análisis, se procede a la ejecución de varios procesos con el objeto de estandarizar, corregir, homologar y/o renombrar la información y de esta manera, obtener el mejor outcome del análisis propuesto. A continuación, se detallan las acciones de limpieza ejecutadas en el archivo:

Tabla 3: ACCIONES DE LIMPIEZA EJECUTADAS

CAMPO	PROBLEMA	ACCIÓN DE SUBSANACIÓN
COD_LOC, COD_PROY, ID_EVALUACION, COD_AUDITOR	no es un campo estandarizado y tiene valores repetidos	incluir una clave primaria alfanumérica
NOMBRE_LOC, POBLACION, OFICINA, PROVINCIA, TITULO_CUESTIONARIO	es un campo que tiene caracteres ascii extendidos y otros signos no reconocidos por utf8	normalización de datos y creación de la base de datos con soporte para utf8mb4_unicode_ci
CP	Los códigos postales no corresponden a la provincia señalada	Homologación de información de códigos postales de España
Fecha de ejecución	fecha en diferentes formatos	normalización de fechas utilizando DateTime y Dataframes

Elaborado por: Grupo L

Procesamiento y Carga: En este punto se carga la información obtenida en el modelo de datos de destino, dejándola disponible para consultas futuras, así como para el desarrollo de la siguiente fase.

Modelos Analíticos & Visualización: Con los insumos obtenidos en las fases anteriores, se procede a analizar la información existente, mediante la creación de distintos modelos analíticos, proyectándolos en herramientas de visualización. Este análisis debe articularse a los objetivos planteados inicialmente.

3. ARQUITECTURA DE REFERENCIA:

Para el desarrollo de este punto, se parte de la primera ley de la arquitectura de software : “Everything in software architecture is a tradeoff”¹; que traducido significa: “Todo en arquitectura de software es contrapartida”. Para este ejercicio, se realiza un trade off entre los 2 tipos principales de arquitectura de referencia, analizando las dimensiones principales bajo el contexto de nuestro proyecto, como se describe a continuación:

Flexibilidad de Cambios en estructura de datos: Se pretende utilizar un enfoque iterativo para la definición del modelo de datos, en este caso exponemos el resultado de la iteración 1 en este documento, por lo que ROLAP al ser más flexible para cambios a nivel de estructura coincide perfectamente con lo que el objetivo planteado.

Complejidad de las consultas y flexibilidad : ROLAP presenta una ventaja al momento de manejo de consultas complejas y debido a que aún no se tiene claro el alcance del análisis del









¹ Richards y Ford, 2020. “Fundamentals of Software Architecture”. O’Reilly Media.

DWH esta característica beneficia al objetivo de la organización.

Frecuencia de cambios en los datos : si bien los datos obtenidos en este momento son estáticos lo por lo que beneficiaría aplicar un modelo MOLAP; uno de los objetivos de este Data Warehouse -DWH- es poder en el futuro, manejar un modelo de ingesta de datos en streaming; por lo tanto, ROLAP es un modelo superior para las necesidades específicas planteadas.

Velocidad de consulta con poco volumen de datos : Aunque la arquitectura ROLAP tiene una desventaja en la velocidad de ejecución frente a MOLAP, para el contexto actual, con la cantidad de datos existentes, este no es un punto de definición.

Ilustración 2: Análisis de la arquitectura de los datos

	ROLAP	MOLAP
Flexibilidad de Cambios en estructura de datos		
Complejidad de las consultas y flexibilidad		
Frecuencia de cambio en los datos		
Velocidad de consulta con poco volumen de datos		

Elaborado por: Grupo L

Finalmente, en la ilustración 2 se observa que ROLAP es la opción más adecuada para el contexto actual de datos y que esta arquitectura se alinea de mejor forma a los objetivos trazados.

4. TECNOLOGÍA OLAP A UTILIZAR:

Toda vez realizado el análisis de factores que inciden en la tecnología a utilizar, se definió el uso de Tableau, por las siguientes razones:

- Experticia del equipo de analítica, procesamiento, análisis multi dimensional y operaciones de Slice and Dice. Adicionalmente, el equipo cuenta con habilidades y experiencia en el manejo de la herramienta Tableau, por lo tanto, el uso de una diferente para la visualización de información involucra un esfuerzo de costo y tiempo posible de evitar.
- Procesamiento y Análisis Multidimensional: Tableau permite a los usuarios analizar datos de manera multidimensional. Se puede organizar y analizar datos en diferentes dimensiones y jerarquías, lo cual es un rasgo clave de los sistemas OLAP.

- Operaciones de Slice and Dice: Tableau permite a los usuarios realizar operaciones de "slice and dice", que significa que pueden ver los datos desde diferentes perspectivas y con distintos niveles de agregación.

5. SI SE UTILIZA ROLAP, ¿CUÁL DE ESTOS DOS MODELOS SE AJUSTARÍA MEJOR: EL MODELO EN ESTRELLA O EL COPO DE NIEVE?

En este apartado, se realiza el análisis de las siguientes dimensiones:

- Complejidad de los Datos: los datos provistos son relativamente simples y no requieren un alto grado de normalización, el modelo en estrella podría ser más adecuado debido a su simplicidad y eficiencia en las consultas.
- Rendimiento de Consultas: Creemos que el "performance" de las consultas a realizar es una prioridad para la ejecución y a pesar de que los análisis no son extremadamente complejos al inicio, el modelo en estrella podría ser preferible.
- Normalización y Detalle de los Datos: Los datos no son complejos, tampoco poseen con múltiples niveles de relaciones entre las dimensiones.

En resumen, la ilustración 3 muestra que el modelo en estrella es la mejor opción bajo el contexto actual por complejidad y rendimiento en las consultas:

Ilustración 3: ANÁLISIS DE MODELOS A UTILIZAR

	Estrella	Copo de Nieve
Complejidad de los datos		
Rendimiento en las consultas		
Normalización y Detalle		

Elaborado por: Grupo L

6. SI SE UTILIZA ROLAP, IDENTIFICAR Y JUSTIFICAR SI EXISTE ALGÚN PROCESO DE DESNORMALIZACIÓN DE INFORMACIÓN QUE SE DEBA REALIZAR

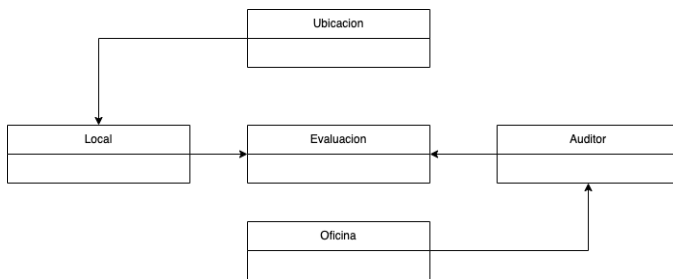
Para el análisis de interpretación de los resultados del procesamiento de datos y obtención de información, se consideró lo siguiente:

- **ÍNDICE ANTIFRAUDE:** los datos de este indicador oscilan entre 0 y 1; donde, 0 significa que no existe riesgo de fraude, mientras que 1 quiere decir que el riesgo es una posibilidad real de ocurrencia.
- El nivel de granularidad a nivel geográfico responde al número de comunidades autónomas existentes en España, ya que es la división administrativa más representativa. Si bien existen localidades ubicadas en el extranjero, se acoge esta clasificación ya que el porcentaje de estas es mínimo.

En resumen no existió un proceso de desnormalización formal, más bien solo uno de normalización que se muestra en los diagramas de base de datos en el punto 11, el script de creación que se lo puede encontrar en este [Archivo](#) y la explicación de carga del ETL en el punto 12 .

7. SI SE UTILIZA ROLAP, INCLUIR UN DISEÑO CONCEPTUAL A MODO EXPLICATIVO JUNTO CON UN DIAGRAMA

Ilustración 4: DISEÑO CONCEPTUAL PLANTEADO



Elaborado por: Grupo L

Entidades :

Ubicación: Esta entidad es la responsable de almacenar las ubicaciones de los locales en donde se realizaron las encuestas de prevención de fraude.

Ubicacion.ID: Este identificador único y numérico fue agregado al momento de la carga de ubicaciones en la base de datos, con el objeto de contar con un identificador único por columna, de esta manera, tener la posibilidad de relacionar los datos con la tabla “Local”.

Ubicacion.Codigo_Postal: es el código postal de la ubicación específica del local.

Ubicacion.Población: se refiere a la población en donde está el local.

Ubicacion.Comunidad_Autonoma: se renombró este campo de población por “Comunidad Autónoma”, para tener un lenguaje ubicuo relacionado al lugar geográfico de donde se obtuvo la

información, que responda a la división administrativa a la que pertenece.

Local:

Local.ID: es la clave primaria de esta entidad, es numérica e incremental, asegura unicidad e identificación en los registros insertados y sirve de relación principal con la tabla de hechos.

Local.Codigo_Local: como se evidenció en el pre-análisis de datos, el Codigo_Local no es un campo homologado por lo que se deduce que viene de diferentes sistemas. No se lo eliminó ya que se considera de utilidad al momento de integrarse con los sistemas de los clientes.

Local.Nombre: es un atributo tipo alfanumérico que contiene información del nombre del local.

Local.Ubicacion_ID: este atributo es la conexión con la entidad de Ubicación ID.

Auditor: Es la entidad responsable de almacenar los datos relacionados a auditor.

Auditor.ID: Es la clave primaria generada para la tabla auditor, cuyo objetivo es el tratamiento y relacionamiento más sencillo entre la tabla de hechos y la tabla auditor.

Auditor.Nombre: Es un campo enriquecido utilizando la librería faker de python. Es el nombre del auditor que aplicó el cuestionario.

Auditor.Oficina: Es el campo relacionado con la tabla de oficina que nos permite saber a qué oficina pertenece el auditor.

Oficina: Es la entidad responsable de almacenar los datos relacionados a Oficina.

Oficina.ID: Es la clave primaria generada para la tabla Oficina. Este campo sirve como un identificador único para cada oficina en la base de datos.

Oficina.Cod_Oficina: Es un campo que almacena el código identificador de la oficina. Puede ser utilizado para fines de clasificación o para relacionarse con sistemas externos que utilicen este código como referencia.

Oficina.Nombre: Este campo almacena el nombre de la oficina. Es útil para identificar de manera legible las diferentes oficinas dentro de la organización.

Evaluación: es la entidad de hechos donde se almacenan las fechas .

Evaluacion.Evaluacion_id: Es la clave primaria de la tabla Evaluación. Representa el identificador único de cada registro de evaluación.

Evaluacion.Resultado: Este campo almacena el resultado de la evaluación. es un campo flotante con un valor entre 0 y uno.

Evaluacion.Titulo: Almacena el título o nombre de la evaluación. Este campo es útil para dar una idea rápida del contenido o el propósito de la evaluación.

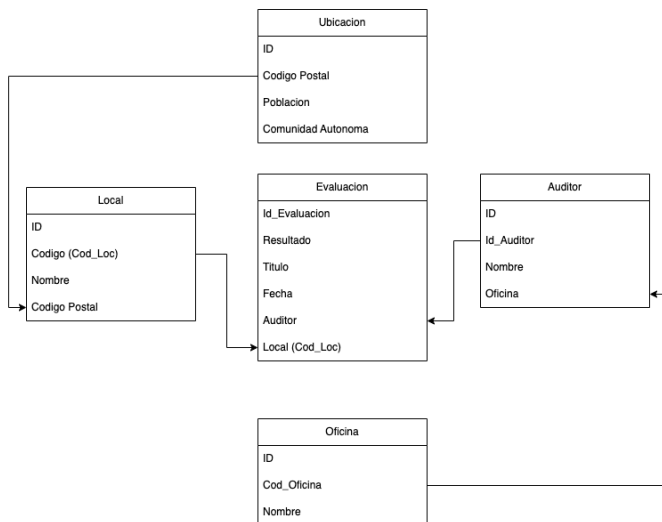
Evaluacion.Fecha: Campo que registra la fecha en que se realizó la evaluación. Es importante para el seguimiento temporal de las evaluaciones.

Evaluacion.Auditor: Es una clave foránea que referencia a Auditor.ID. Indica qué auditor realizó la evaluación.

Evaluacion.Local: Es una clave foránea que referencia a Local.ID Este campo vincula la evaluación con el lugar específico donde se realizó.

8. SI SE UTILIZA ROLAP, INCLUIR UN DISEÑO MODELO LÓGICO:

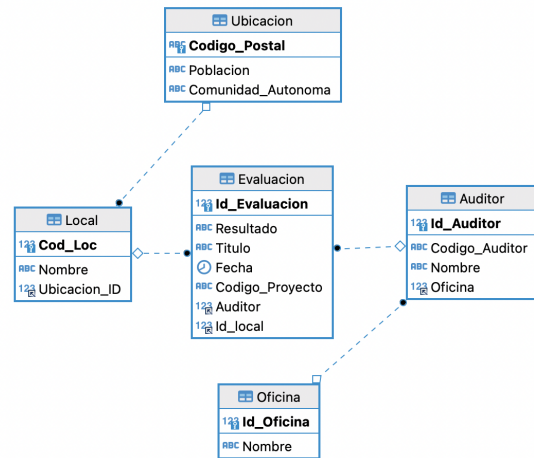
Ilustración 5: DISEÑO DEL MODELO LÓGICO PLANTEADO



Elaborado por: Grupo L

9. SI SE UTILIZA ROLAP, INCLUIR UN DISEÑO MODELO FÍSICO:

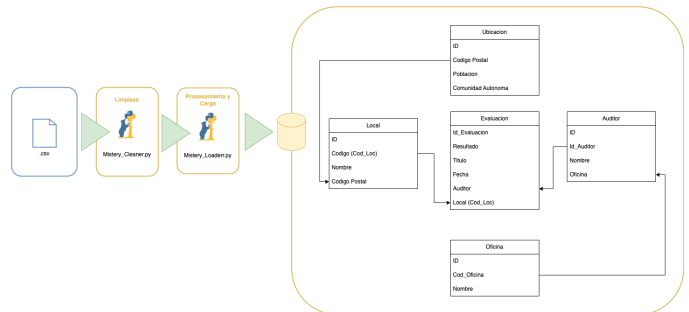
Ilustración 6: DISEÑO DEL MODELO FÍSICO PLANTEADO



Elaborado por: Grupo L

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 O TRABAJO GLOBAL “Global_IMF.kjb”

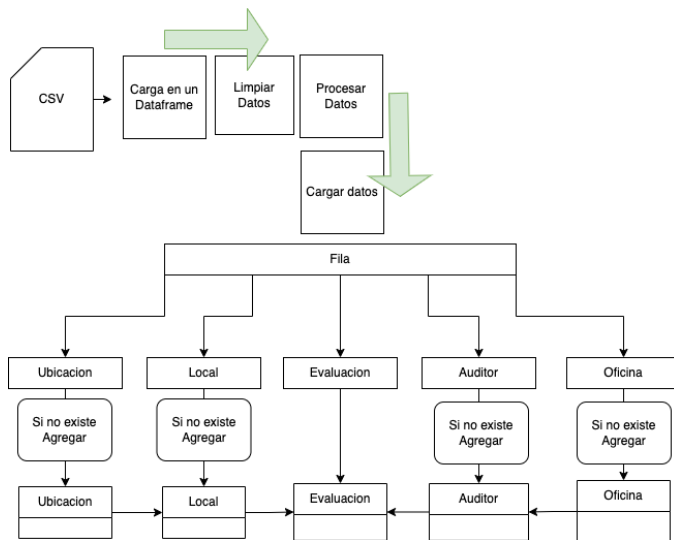
Ilustración 7: IMPLEMENTACIÓN DEL PROCESO ETL



Elaborado por: Grupo L

El Archivo `MysteryLoader/MysteryDaemon.py` es el encargado de tomar, limpiar, procesar y cargar los datos en la BD de MySQL. La conexión está manejada por `sqlAlchemy` y cuenta con algoritmos para agregar los registros faltantes, evitando duplicar información en las tablas del modelo de BDD. Por ejemplo, en la tabla oficina sin este algoritmo podríamos tener 32797 registros, pero actualmente obtenemos 13.

Ilustración 8: FLUJO DE FUNCIONAMIENTO DEL ETL

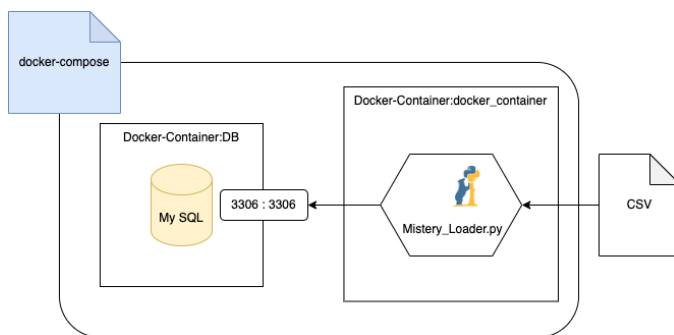


Elaborado por: Grupo L

11. IMPLEMENTACIÓN DEL MODELO MULTIDIMENSIONAL DISEÑO MEDIANTE LOS PUNTOS ANTERIORES. SE DEBE REALIZAR CON LA HERRAMIENTA WIZARD FACILITADA Y MOSTRADA EN VIDEOS ANTERIORES.

La implementación se realizó mediante el uso de tecnología Docker y Docker Compose. Su principal ventaja es que puede ser transportado a cualquier nube que soporte tecnología de contenedores sin mayores dificultades para su configuración, como lo muestra la siguiente ilustración:

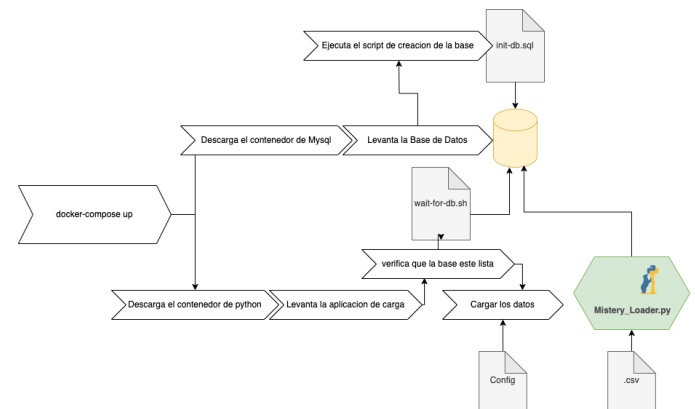
Ilustración 9: IMPLEMENTACIÓN ETL Y DWH UTILIZANDO INFRAESTRUCTURA COMO CÓDIGO



Elaborado por: Grupo L

Para el levantamiento del datawarehouse únicamente se debe ingresar a la solución en la carpeta “Infraestructura” y ejecutar: docker-compose.up. La secuencia de levantamiento programada en el archivo es la siguiente:

Ilustración 10: FLUJO DE CONSTRUCCIÓN DE LA INFRAESTRUCTURA COMO CÓDIGO



Elaborado por: Grupo L

Como lo muestra la ilustración 10, al ejecutar el comando señalado, Docker levanta los 2 contenedores asociados al proceso: el contenedor de la base de datos y el contenedor del ETL. Se levanta el contenedor de bd con MySQL, y automáticamente ejecuta el archivo de la base de datos : init-db.sql creando la estructura del DWH.

A la par, en el contenedor del ETL se está ejecutando el archivo [wait-for-db.sh](#) para comprobar que la base está preparada para recibir los datos. Una vez lista, la aplicación de carga, insertará los datos en la base; aquí se puede indicar el número de registros a leer del csv en el archivo config.yaml.

12. ANÁLISIS DEL MODELO. SE SOLICITA REALIZAR AL MENOS, UN ANÁLISIS HACIENDO USO DE UN MODELO MULTIDIMENSIONAL QUE REFLEJE ALGUNA SITUACIÓN RELEVANTE DE SER EXPLICADA:

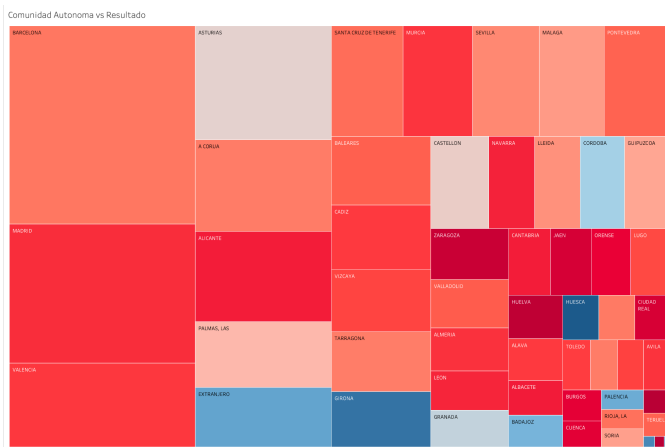
La empresa Mystery Shopping ha establecido tres KPIs (Indicadores Clave de Rendimiento) fundamentales para evaluar aspectos críticos de sus operaciones y servicios. El primer KPI es particularmente innovador, enfocado en calcular un índice antifraude para empresas clientes. Este índice mide la eficacia de las medidas preventivas contra el fraude implementadas en las empresas clientes, proporcionando una valoración cuantitativa de su capacidad para detectar, prevenir y responder a posibles actividades fraudulentas. Este KPI no solo refleja la solidez de las prácticas de seguridad de las empresas clientes, sino que también destaca el valor añadido que la empresa aporta a través de sus servicios de consultoría y asesoramiento en seguridad y prevención de fraude.

El segundo KPI mide el número de encuestas realizadas por auditor, lo cual es esencial para evaluar la productividad y la eficiencia operativa del personal de auditoría. Este indicador ayuda a asegurar que los recursos de auditoría se utilicen de manera óptima y que cada auditor esté contribuyendo efectivamente a la recopilación de datos importantes.

Finalmente, el tercer KPI se centra en las encuestas realizadas en una fecha específica, permitiendo un seguimiento detallado de las operaciones diarias. Este KPI es crucial para monitorear la coherencia y el ritmo de trabajo, asegurando que se cumplan los plazos y se mantenga un alto nivel de rendimiento en todas las actividades relacionadas con las encuestas.

En conjunto, estos tres KPIs proporcionan una perspectiva integral y multidimensional del rendimiento y la eficacia, lo que facilita la toma de decisiones estratégicas y el fortalecimiento de las relaciones con los clientes.

Ilustración 11: Comunidad autónoma vs resultado



Elaborado por: Grupo L

La gráfica corresponde a un mapa de calor o treemap, en donde cada bloque representa una Comunidad Autónoma de España, y podemos inferir que:

- **Cantidad de Encuestas:** El tamaño de cada bloque sugiere la cantidad de encuestas realizadas en cada Comunidad Autónoma. Las áreas con bloques más grandes, como Madrid y Barcelona, parecen tener un número mayor de encuestas realizadas, lo que podría indicar que estas áreas son de especial interés para la empresa o que tienen una mayor población y, por lo tanto, son más propensas a ser seleccionadas para encuestas.
- **Probabilidad de Fraude:** La intensidad del color rojo en cada bloque indica la probabilidad de fraude asociada con cada Comunidad Autónoma. Los bloques más oscuros, que muestran un rojo más intenso, sugieren una mayor probabilidad de fraude en las encuestas realizadas en esas áreas. Esto puede ser una señal para que la empresa aumente sus controles de prevención de fraude o investigue más a fondo las causas de esta tendencia.

Áreas de Preocupación: Las Comunidades Autónomas con bloques grandes y de color rojo intenso serían de particular preocupación ya que combinan un alto volumen de encuestas con una alta probabilidad de fraude. Por ejemplo, si Madrid tiene un bloque grande y muy rojo, sería un área clave para centrar los esfuerzos de mitigación de riesgos.

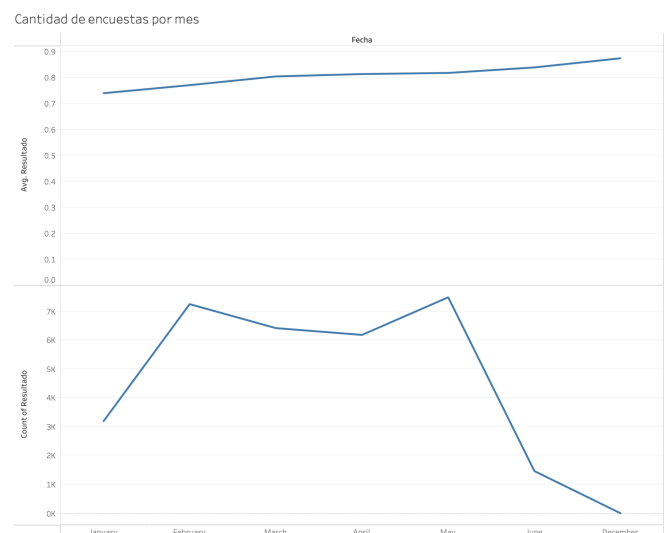
Áreas de Menor Riesgo: Por otro lado, las áreas con bloques más claros (menos intensidad de rojo) y de menor tamaño, como Asturias en la visualización, podrían indicar un menor número de encuestas y una menor probabilidad de fraude, lo que las convierte en áreas de menor preocupación en términos de fraude.

Distribución Geográfica: La distribución de los bloques y sus colores también podría proporcionar insights sobre la geografía del fraude en las encuestas, lo que podría indicar tendencias regionales o culturales que la empresa podría querer investigar.

Extranjero: El bloque etiquetado como "Extranjero", de un color azul diferente, podría indicar datos de encuestas realizadas fuera de España, lo que sugiere que la empresa también opera en mercados internacionales. La diferencia en el color puede indicar una categorización distinta para el análisis, como un sistema de control o riesgo de fraude diferente.

Es importante mencionar que para una interpretación precisa y detallada se requiere acceso a los datos subyacentes y un entendimiento completo del contexto empresarial.

Ilustración 12: AVG Resultado vs. fecha y Número de Encuestas vs. Fecha



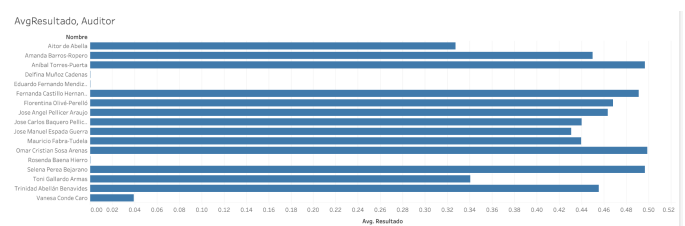
Elaborado por: Grupo L

Tendencia de la Probabilidad de Fraude: La línea superior, que representa la probabilidad promedio de fraude (Avg

Ilustración 13: Numero de encuestas mayor a la media vs Auditor



Ilustración 14: Auditores con índice menor a la media



Elaborado por: Grupo L

13 . Conclusiones

- La adopción de IaC (infraestructura como código) y Docker en la implementación de un Data Warehouse es una estrategia que promete agilidad, escalabilidad y consistencia en la gestión de la infraestructura de datos. IaC permite la automatización del aprovisionamiento y la gestión de la infraestructura a

través de código, lo cual reduce el potencial de errores humanos y acelera significativamente el tiempo de despliegue de los recursos necesarios. Además, asegura que la infraestructura sea reproducible y fácilmente versionable, lo que simplifica las actualizaciones y la mantenibilidad.

- Implementar el modelo estrella en MySQL para la construcción de un Data Warehouse y utilizar Python para su alimentación y manipulación de datos es una combinación poderosa que aprovecha la robustez de un sistema de gestión de bases de datos relacional con la flexibilidad y potencia de un lenguaje de programación de alto nivel.
- La arquitectura del modelo estrella favorece la simplicidad y la eficiencia en las consultas analíticas, ya que las tablas de dimensiones se organizan alrededor de una tabla de hechos central, lo que facilita la comprensión y el acceso a los datos. La combinación del modelo estrella en MySQL con la carga y manipulación de datos a través de Python es una estrategia efectiva para el desarrollo de un Data Warehouse. Ofrece una solución económica, flexible y potente que puede ser personalizada para satisfacer las necesidades específicas de análisis de datos de una organización, aprovechando al mismo tiempo las tecnologías de código abierto y las comunidades activas que las respaldan.