Universidad Técnica Nacional
Ingeniería del Software
Minería de Datos
ISW-911

Proyecto #1
Data Warehouse

Documento Escrito

Estudiantes Encargados

Eduardo Fabricio Cordero Córdoba
Neuman Josafath Ramírez Quesada
Benjamín Gadiel Sandí Salas
Roney Alfonso Valdelomar López

Profesor
Freddy Gerardo Rocha Boza

III Cuatrimestre
Año 2024

Tabla de Contenidos

Introducción

El presente documento busca dar una solución al problema presentado en el apartado asignado, utilizando una selecta variedad de herramientas y métodos informáticos para la creación de una base de datos con un mercado de datos (warehouse) útiles para los propósitos de la empresa, los cuales buscan analizar los periodos con mayor actividad en ventas y que les ayude en una mejor planificación en sus estrategias de negocio.

Para dar solución al problema planteado, se utilizaran herramientas de software como SQL Server Management Studio, Oracle SQL Manager, Visual Code, Visual Studio 2015, bases de datos como Oracle y SQL y lenguajes de programación como Python.

Se espera que la solución dada sea acertada y eficaz para la solución del problema.

Problema

La empresa Technologies Inc S.A, tiene la necesidad de llevar a cabo la implementación de un data warehouse. Se organizará un grupo de personas para llevar a cabo una investigación teórico-práctica centrada en el proceso de ETL (Extracción, Transformación y Carga) y Data Warehouse. Esto con el fin de brindarle a la empresa información importante acerca sus estrategias de ventas actuales, nueva información de posibles estrategias de ventas, posibles mejoras en estrategias o predicciones de consumo a futuro, en busca de un crecimiento empresarial ante las nuevas tendencias del mercado. Permitiendo así al grupo obtener habilidades competitivas y conocimientos sólidos en el ámbito del data warehousing que les permitirá brindar un mejor apoyo a la empresa y mejorar su rendimiento personal como trabajadores.

Objetivo General

- ➔ Crear un Data Warehouse que permita realizar un análisis detallado de los periodos de mayor y menor actividad de ventas, con el fin de optimizar la planificación de inventario y mejorar la toma de decisiones en relación con la estrategia comercial.

Objetivos Específicos

- ➔ Desarrollar un sistema para el almacenamiento, procesamiento y análisis de datos de ventas.
- ➔ Crear reportes periódicos para identificar periodos de mayor y menor actividad de ventas.
- ➔ Proveer datos segmentados de clientes y rendimiento de empleados para optimizar decisiones en marketing y recursos humanos.

Descripción de la solución

1. Requisitos del Negocio:

1. **Análisis del comportamiento de ventas**: La empresa necesita identificar los periodos de mayor y menor actividad de ventas para optimizar la estrategia de inventario y ventas, basándose en patrones históricos.
2. **Segmentación de clientes**: Se espera que la empresa pueda agrupar a los clientes en diferentes segmentos basados en su comportamiento de compra (frecuentes, esporádicos, de alto valor) para mejorar el enfoque de las campañas de marketing.
3. **Evaluación de desempeño de empleados y oficinas**: Analizar el rendimiento de los empleados y oficinas, identificando a los mejores vendedores y oficinas más productivas, permitiendo así ajustar las políticas de incentivos y mejorar la productividad.

2. Necesidades del usuario

Usuarios Finales:

- **Gerentes de Ventas**: Necesitan identificar los periodos de mayor y menor actividad comercial para ajustar las estrategias de inventario y personal.
- **Marketing**: Necesitan segmentar a los clientes por comportamiento de compra para diseñar campañas específicas que maximicen las ventas.
- **Recursos Humanos**: Requieren evaluar el desempeño de los empleados de ventas para tomar decisiones sobre incentivos, promociones, o áreas de mejora.

Necesidades específicas:

- **Gerentes de Ventas**:
  - Visualización de patrones históricos de ventas para identificar picos y caídas.
  - Acceso a informes que permitan planificar inventario para satisfacer la demanda en los periodos de mayor actividad.
- **Marketing**:
  - Acceso a datos segmentados de los clientes para diseñar campañas personalizadas.
  - Informes sobre clientes de alto valor, frecuentes, y esporádicos para mejorar la retención y optimizar estrategias de marketing.
- **Recursos Humanos**:
  - Informes sobre la productividad de los empleados y rendimiento de oficinas.
  - Datos que permitan identificar empleados con alto rendimiento y aquellos que necesitan apoyo para mejorar.

3. Objetivos del DW

El Data Warehouse (DW) debe cumplir con los siguientes objetivos claros y medibles para satisfacer las necesidades del negocio:

Objetivo Principal:

- ➔ **Análisis de los periodos de mayor y menor actividad de ventas:** El objetivo principal de este proyecto es realizar un análisis exhaustivo de los datos históricos de ventas para identificar patrones de actividad comercial a lo largo del tiempo, con el fin de reconocer los periodos de mayor y menor volumen de ventas. Este análisis abarca la identificación de los meses o trimestres con más alta y baja actividad de ventas en los últimos cinco años, facilitando una visión detallada que permitirá tomar decisiones más informadas. Con estos datos, la organización podrá optimizar la planificación de inventarios, asegurando un abastecimiento adecuado durante los periodos de alta demanda, y podrá ajustar sus estrategias de marketing y ventas en los periodos de menor actividad. La finalidad es maximizar el rendimiento operativo y mejorar la eficiencia en el uso de recursos, aprovechando los datos de ventas como base para decisiones estratégicas.

Objetivos Secundarios:

1. **Segmentación de clientes por comportamiento de compra**:
   Este objetivo busca analizar y agrupar a los clientes en diferentes segmentos según su historial de compras, tales como clientes “frecuentes”, “esporádicos” y “de alto valor”, entre otros. Esta segmentación proporcionará al equipo de marketing una comprensión más detallada del perfil de cada cliente, permitiendo diseñar campañas más específicas y personalizadas, enfocadas en mejorar la retención y aumentar las ventas dirigidas. El análisis incluirá la generación de informes mensuales que clasifiquen a los clientes en estos grupos de valor, con el propósito de mejorar las tasas de retención de clientes en al menos un 10% y optimizar los resultados de campañas enfocadas, apoyando así el logro de una relación más sólida con los clientes y la maximización de ingresos.

2. **Evaluación del rendimiento por empleado o por oficina**:
   Este objetivo se enfoca en analizar el desempeño de los empleados y de las oficinas de ventas de la empresa, con el propósito de identificar a los empleados más productivos y las oficinas con mejor rendimiento. El análisis se realizará a través de informes trimestrales detallados sobre el volumen de ventas alcanzado por cada empleado y oficina, lo que permitirá a la organización implementar ajustes en los incentivos y en los procesos laborales. Con esta información, la empresa podrá mejorar la productividad general de los equipos de ventas, orientando esfuerzos y recursos hacia las prácticas más efectivas y, al mismo tiempo, fomentando un ambiente de reconocimiento y motivación que impulse el crecimiento continuo y el compromiso del personal.

4. Fuentes de Datos

Para alimentar el Data Warehouse y cumplir con los objetivos descritos, es necesario identificar y consolidar las diferentes fuentes de datos que serán utilizadas. Estas fuentes pueden ser tanto internas como externas.

Fuentes de Datos Internas:

1. **Datos de ventas**:
   - Detalles de cada transacción (fecha, cliente, empleado, producto, cantidad, precio, descuento, etc.).
   - Historial de ventas de los últimos años.

2. **Datos de empleados**:
   - Base de datos del sistema de recursos humanos con información sobre el desempeño y las ventas gestionadas por cada empleado.

3. **Datos de clientes**:
   - Historial de compras de clientes, comportamiento de compra, segmentos demográficos, etc.

4. **Datos de productos**:
   - Información detallada sobre los productos vendidos, como categorías, precios, fechas de lanzamiento, etc.

Mapeo de Tablas de Datos:

A continuación, se muestra cómo se utilizarán las diferentes tablas de dimensiones y hechos en el Data Warehouse:

- **Tabla de hechos**: FactVentas (almacena las transacciones de ventas).

- **Dimensiones**:
  - DimCliente (contiene información detallada sobre los clientes).
  - DimEmpleado (información sobre los empleados de ventas).
  - DimProducto (detalles de los productos vendidos).
  - DimFecha (dimensión temporal para analizar las ventas a lo largo del tiempo).

5. Calidad de Datos

Los datos extraídos de las base de datos fuentes permanecían en dos diferentes idiomas, unos en ingles y otros en español, además de que han datos simplificados que se tenían que extender como ejemplo está el caso de los datos que se refieren al país de un cliente, ya que estos habían unos en ingles, en español y otros bajo el acrónimo del país.

Por lo que se tuvo que emplear herramientas de terceros para la traducción de los datos, todos sean bajo un mismo lenguaje, tamaño de letra y puedan coexistir en la mismas columnas, además de añadir nuevos parámetros de traducción de datos para algunos datos bajo un acrónimo, por ejemplo, se añadió el parámetro “UR” para que lo traduzca a “United Kingdom”, así con algunos datos más, donde se mostrada en la sección de “Scripts Documentados”.

6. Granularidad

El nivel de detalle en cada dato será fino, guardando cada transacción realizada, su tiempo de venta, vendedor, cliente, además de guardarse los datos de forma descriptiva para evitar confusiones por acrónimos, todo en mayúsculas y a un único lengua, español para casos de este proyecto, mientras que el manejo de la base de datos, nombramiento de tablas o columnas, sera en ingles.

7. Dimensiones y Medidas

Las dimensiones y medidas de los datos a trabajar serán:

Tabla	Dimensión	Medida
Clientes	Nombre Compañía	Texto
	Nombre Contacto	Texto
	País	Texto
Empleados	Nombre Completo	Texto
	Título/Cargo	Texto
	País	Texto
	Reporta A	Texto
Productos	Nombre Producto	Texto
	Categoría	Texto
	Precio por Unidad	Numérico
	Unidad en Stock	Numérico
	Cantidad por Unidad	Texto
Fechas	Año	Numérico
	Mes	Numérico
	Día	Numérico
	Cuatrimestre	Numérico
	Semestre	Numérico
	Semana	Numérico
	Día de la Semana	Texto
	Día del Año	Numérico

8. Proceso ETL

Para la extracción de datos de las bases de datos, se creó una base de datos “transitoria”, en SQL (nombrada staging para este proyecto), donde poner todos los datos en un solo sitio y poder realizar modificaciones o limpieza, esto con el fin de no modificar las bases de datos fuentes, además de tener un lugar donde empezar con el análisis de cuáles tablas y datos se necesitan, en base al valor de los datos y la importancias de estos para el análisis de otros datos.

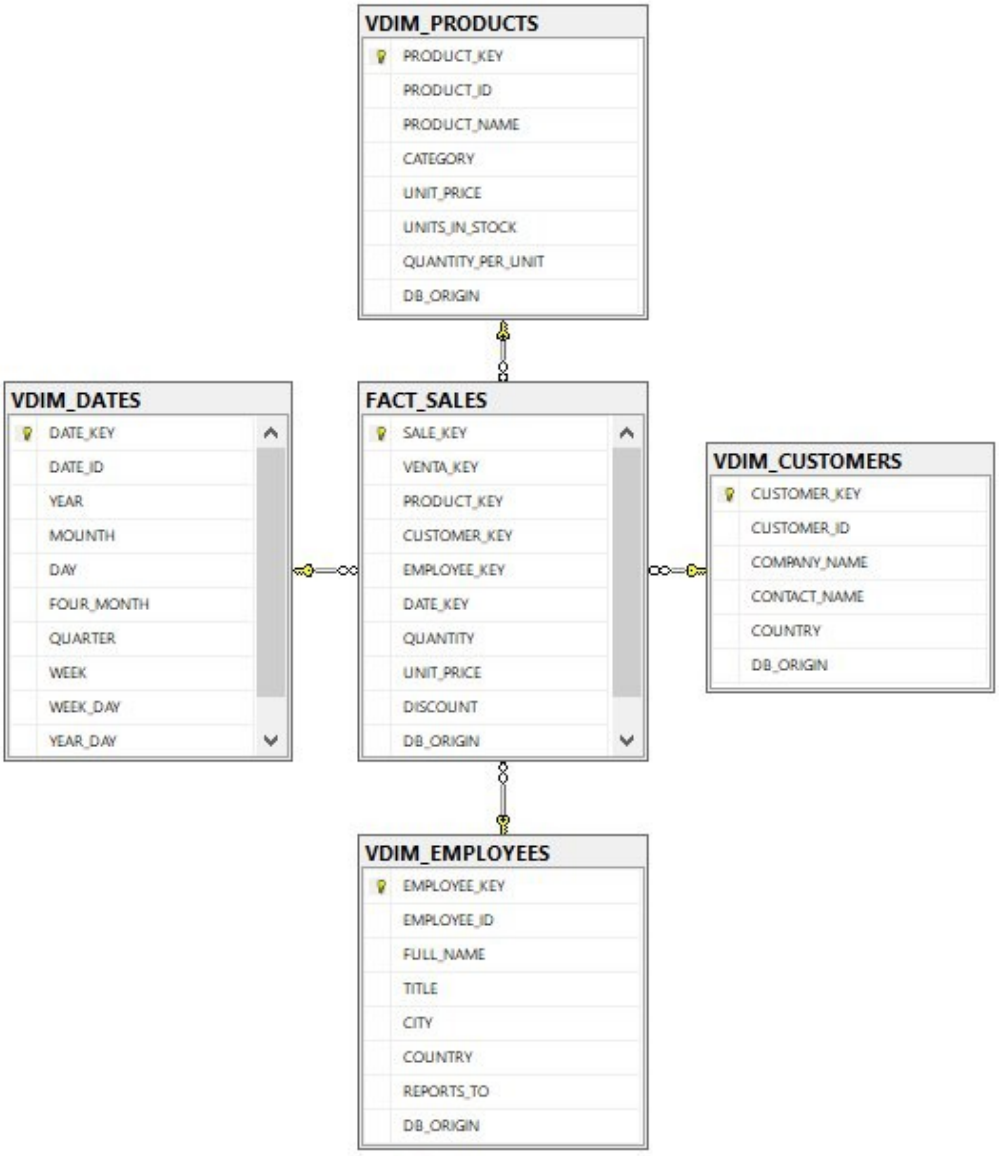
Con los datos dentro de la base de datos transitoria, se empieza el análisis de los datos y cuáles columnas proporcionan datos relevantes; Algunas columnas, con posibles datos importantes o de valor para otros datos, se tuvieron que descartar por la gran ausencia de los datos específicos que van en dicha columna, aunque en algunos casos, columnas con datos nulos pudieron ser rellenadas con datos verídicos, gracias a un análisis hecho junto a los otros datos que se poseía del mismo conjunto.

Una vez con el análisis hecho de cuales datos se van a utilizar, donde se deben hacer la transformación de datos y se define como se van a manejar los datos dentro del warehouse, comienza la creación de las vistas dentro de la base de datos transitoria, donde van a tener los datos necesarios y útiles que pueden ayudar a la empresa en su problemática.

Con los datos en las vistas ya listos, se empieza la transformación todos los datos en base a las reglas puestas, en este caso: todo mayúsculas para una visualización más limpia de los mismos, se mantuvieron de forma descriptiva y simples para evitar confundirlos con otros acrónimos, además de mantenerlos todo a un mismo idioma, para este caso español; A estos también se le añadió una nueva columna con el nombre de la base de datos de origen, esto para luego reconocer de cuál base de datos fuente llegó cada dato y poder tener una trazabilidad de los mismos.

Por último, con los datos ya transformados, limpiados, solucionados aquellos datos nulos y formateados, estos fueron pasados al warehouse donde se les dio un nuevo identificador incremental que ayude a enlazarlos.

Diagrama E/R del DW



Scripts Adjuntos (Documentados)

Todos los scripts utilizados para dar solución a la problemática planteada en este proyecto estarán en una carpeta o archivo comprimido .zip, junto a este documento escrito, llamado “SCRIPT-MINERIA” junto a los comentarios de cada uno, en este apartado se hablara a un nivel más superficial sobre estos archivos adjuntos.

Se utilizó el lenguaje de programación Python para el desarrollo de los archivos para el ETL, junto a un archivo .env para los valores globales que serán utilizados durante la ejecución de algunos scripts, algunos ejemplos de valores globales que se necesitaron para este caso fueron los nombres de las diferentes bases de datos donde se extrae la información necesaria para el warehouse, el nombre de la base de datos transitoria y el nombre de la base de datos warehouse que almacenaría los datos, también se tuvo que añadir las credenciales de acceso de los usuarios asociados a las bases de datos en SQL y Oracle, así como otros datos necesarios para el correcto funcionamiento de los demás scripts.

Para empezar a utilizar los scripts de python, desde una consola tiene que ejecutar el archivo “menu.py” el cual importa los archivos necesarios y ejecuta un pequeño menú interactivo, el cual permite al usuario cargar los datos a una base de datos transitoria, crear las vistas con los datos necesarios para transformar, y migra los datos hacia la base de datos warehouse, aplicando el proceso de transformación mientras realiza la migración, de una forma segura y confiable.

Todos los archivos ejecutados en el “menu.py” están dentro de las carpetas “ETL”; En “SQL” se guardan archivos .sql con las diferentes consultas SQL que se utilizaron en el proceso ETL para la extracción, modificación y carga de los datos de una base de datos a otra. Estas consultas pueden ser usadas para la creación de las vistas con los datos necesarios para el warehouse y la creación de las tablas dentro del warehouse desde una herramienta, como lo es SMSS (SQL Management Server Studio).

Dentro de la carpeta “ETL” se encuentra más carpetas como “conexiones”, la cual contiene los archivos para realizar las conexiones a las dos diferentes base de datos, Oracle y SQL, además de contener el patrón Singleton para mantener la instancia de conexión a las bases de datos.

Dentro de la carpeta “config” se encuentra archivos .json, funcionales para la transformación de los datos, creación de las vistas en las bases de datos y de las tablas dentro del warehouse, ya que estos también contienen los datos de las columnas que deben ser transformadas, traducidas y modificadas para su correcto uso dentro del warehouse.

Luego estaría la carpeta “scripts” con los archivos .py, donde tendríamos el archivo con las operaciones para extraer los datos de las bases de datos fuentes; También se encuentra el

archivo con las operaciones necesarias para la creación de las tablas en el warehouse, así como otro más para exportar los datos de la base de datos transitoria a la base de datos warehouse, y un último archivo con la funciones para la transformación de los datos dentro de las vistas.

Por último esta la carpeta “utilidades”, la cual contiene archivos con funciones más concretas y de uso concurrentes en otros procesos dentro de la carpeta “scripts”, como pueden serlo el análisis de tablas o datos, ejecución de algunas funciones SQL, formateo, transformación de datos, manejo de los datos, creación de llaves foráneas o traducción de los datos migrados a la base de datos warehouse.

Es bueno destacar que para información más específica sobre la funcionalidad de cada archivo, cada uno viene documentado más en detalle y sobre la funcionalidad de cada una de las funciones dentro de cada archivo y para qué funciona cada uno, además de las librerías que requiere cada uno o que otros archivos utiliza cada uno.

Conclusión

El proyecto se vuelve muy interesante y permite ver todo el proceso que debe realizar una empresa para poder hacer un adecuado análisis de los que ellos mismo pueden estar generando gracias a las bases de datos que utilizan. Además se pudo apreciar cómo utilizar diversas herramientas, que comúnmente no se ven trabajar adecuadamente entre sí, realizar diferentes tareas para resolver un problema de forma eficaz y segura.

Es apreciable el esfuerzo que se ve desde inicio a fin, desde la extracción de los datos de las diferentes bases de datos fuentes de información, la selección de los datos ya cuando se agrupan todos en una sola base de datos transitoria, el análisis de que se realiza sobre los datos en base al objetivo del warehouse y reglas del negocio, hasta la transformación de los datos y migración de los datos necesarios al warehouse.

Ya con los datos montados en el mercado de datos, listos para ser usados, es posible manipularlos mediante herramientas de análisis de datos, desde la más sencilla como Excel, hasta la más extravagante como Power BI, permiten que los datos puedan ser vistos de una forma graficas hacia el cliente y puedan realizarse predicciones o modificaciones a planes de negocio o que la empresa pueda implementar una nueva estrategia de negocio, gracias a adecuada estrategia para tratar los datos y transformarlos.