Informe Final – Diseño de Arquitectura de Datos Moderna

1. Introducción

Este informe presenta el diseño de una arquitectura de datos moderna para una organización de la industria del retail, utilizando como base los conjuntos de datos proporcionados sobre clientes, productos y ventas. El objetivo principal es transformar datos sin normalizar en información valiosa que permita tomar decisiones estratégicas, implementando los conceptos clave aprendidos en el módulo como: Data Lake, Data Warehouse, Data Marts y procesos ETL.

La solución propuesta aborda el desafío de integrar fuentes de datos dispersas en un flujo coherente que va desde la ingesta inicial hasta el análisis final de los datos, cumpliendo con los criterios solicitados.

2. Diseño de la arquitectura

Presentamos la siguiente arquitectura propuesta que contiene el flujo desde el origen que está compuesto por los archivos de clientes, productos y ventas.

La carga e ingesta de datos se realizará a través de los archivos mencionados anteriormente que contienen la información fundamental del negocio.

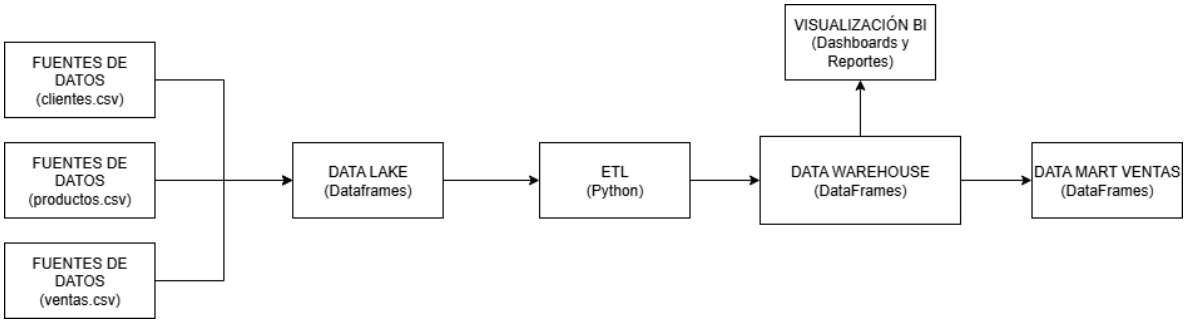
Esta ingesta se transforma en un **Data Lake** que almacena una copia exacta de los datos de origen antes de cualquier modificación para fomentar la transparencia en la obtención de información desde los datos brutos.

Transformación de datos usando Python para limpieza, validación, unión de tablas y cálculos, este es el proceso de **ETL** el cual realizaremos en Python para evaluar el contenido adecuado evaluar los campos y su potencial a ser usado en reportes.

Generar un esquema dimensional con tablas de dimensiones y hechos almacenados como DataFrames representando la sección de **Data Warehouse** en la imagen a continuación con la información depurada, con foco a la maximización de contenido para la preparación de reportes, visualización por dashboards como es Power BI.

DataFrames especializados exportados desde el Data Warehouse para análisis específicos.

Con el objetivo de entregar una solución integral para los diversos equipos de la compañía representada por el **Datamart** de ventas.



3. Procesos ETL y almacenamiento

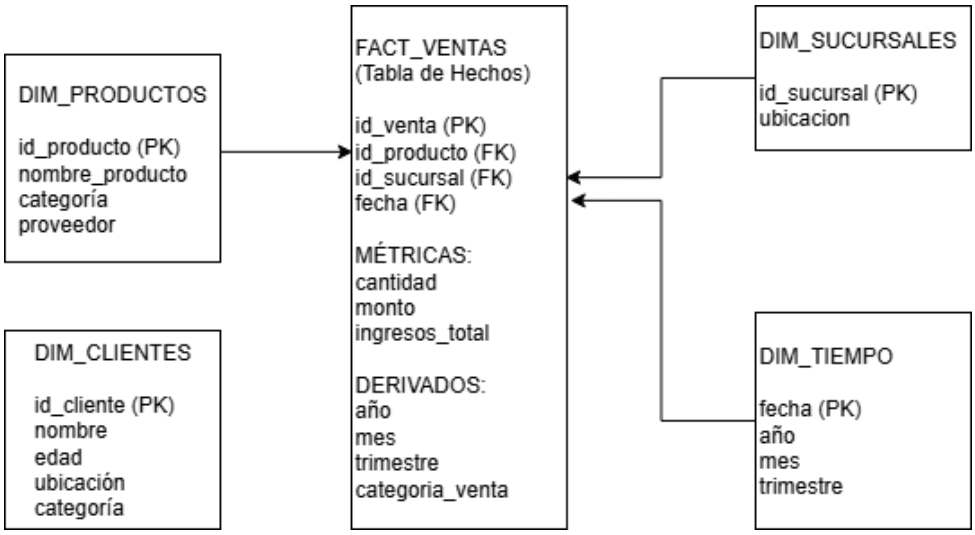
Para realizar el proceso de ETL utilizamos python el cual puede visualizar el código en el siguiente link [Colab ETL Python] revisando la existencia de datos nulos y duplicados.

Generamos copias del data lake para generar el proceso de transformación, modificando fechas, calculando métricas como es el caso del ingreso total.

Generamos una tabla de hechos integrada que contiene el compilado de la información de las tablas y por último agregar una categoría de análisis .

4. Modelo multidimensional

Consideramos modelo “Estrella”:



Con esta información podemos obtener el Data marts

1. Ventas por categoría de producto:
   - Dimensión: tiempo, categoría producto
   - Hechos: monto total, cantidad de ventas
2. Rendimiento por sucursal:
   - Dimensión: tiempo, ubicación
   - Hechos: volumen de ventas, productos más vendidos

5. Evaluación de calidad de datos

Las métricas de Calidad Evaluadas fueron:

Completitud:

- Clientes: 0 valores nulos detectados
- Productos: 0 valores nulos detectados
- Ventas: 0 valores nulos detectados
- Completitud general: 100%

Unicidad:

- Clientes: 0 registros duplicados
- Productos: 0 registros duplicados
- Ventas: 0 registros duplicados
- Sin problemas de duplicación

Consistencia:

- Rango temporal coherente en todas las transacciones
- Valores numéricos dentro de rangos esperados
- Integridad referencial verificada entre tablas

Validez:

- Formatos de fecha correctos y parseables
- Valores monetarios y cantidades positivos
- Códigos de productos válidos en todas las ventas

Lo que en resumen nos entrega la siguiente tabla

Critierio	Resultado	Problemas Detectados
Completitud	100%	Ninguno
Consistencia	100%	rmatos, categorías y valores coherentes
Exactitud	100%	idaciones de rango y formato correctas
Unicidad	100%	Sin duplicados

Esta validación nos permite visualizar la calidad de los datos antes del procesamiento ETL, asegurando que solo datos limpios ingresen al Data Warehouse.

6. Visualización y análisis

Para esta sección puede visualizar diversos dashboard creados con el apoyo de la información del data warehouse en el Colab dashboards.

Dashboard 1: Análisis Clave de Ventas

1. Evolución de Ingresos Mensual

   Gráfico de líneas que muestra tendencias temporales de ingresos a lo largo del período analizado

2. Top 5 Productos Vendidos

   Gráfico de barras que presenta el ranking por cantidad vendida de los productos más exitosos

3. Ingresos por Categoría

   Gráfico de torta que muestra la distribución porcentual del revenue por segmento de productos

4. Ingresos por Sucursal

   Gráfico de barras comparativo del performance entre diferentes ubicaciones

Dashboard 2: Reportes de Análisis Avanzado

1. Tendencia de Ingresos y Cantidad

   Gráfico de doble eje que combina evolución temporal de ingresos y volúmenes de venta

2. Eficiencia por Sucursal

   Scatter plot que relaciona cantidad promedio vs. ingresos promedio por sucursal

3. Análisis Comparativo por Categoría

   Gráfico dual que combina ingresos totales con número de transacciones por categoría

4. Distribución de Ticket Promedio

   Histograma que muestra la distribución del valor de las transacciones individuales

Reporte Ejecutivo de KPIs

Se desarrolló una tabla ejecutiva consolidada con los indicadores clave de rendimiento:

- Total de Ingresos
- Ingreso Promedio
- Total de Unidades
- Número de Transacciones
- Productos Únicos:
- Sucursales Activas
- Período de Análisis

7. Conclusiones y recomendaciones

Este ejercicio confirma que el valor real de los datos emerge al integrar tres pilares:

1. Arquitectura escalable,
2. Procesos ETL robustos
3. Modelos analíticos orientados al negocio.

La lección más relevante es que incluso con datos limitados (como los CSV iniciales), una estrategia bien diseñada puede revelar oportunidades ocultas y guiar decisiones con impacto tangible en ventas y eficiencia operativa.

Recomendaciones:

- Eficiencia ETL: Python demostró ser efectivo para transformar y limpiar los datos, por lo tanto se recomienda como herramienta para el análisis de datos comerciales
- Incorporar campo no mapeable: actualmente la información del cliente no se encuentra disponible para ser mapeada e incorporar en el análisis, es por esto que se recomienda incorporar el campo a través de un desarrollo, de esta manera en el futuro poder identificar comportamiento de los clientes en particular si este visita más de una sucursal por ejemplo.
- Segmentación: Los clientes de Valparaíso aunque representaban solo el 22% del total son los generaron el 35% de las ventas, destacando la importancia de estrategias de fidelización por mencionar un ejemplo.
- Enfoque regional: Valparaíso y Santiago concentraron el mayor volumen de ventas, se sugiere priorizar recursos en estas zonas.
- Productos Clave: Tecnología y Alimentos, lideraron las ventas, mientras que Ropa tuvo menor desempeño, lo que podría orientar el inventario.
- Data Marts útiles: Las visitas multidimensionales agilizaron el análisis sin sobrecargar el Data Warehouse.
- Calidad de datos: no se detectaron inconsistencias en los registros, aun así, es necesario recalcar la necesidad de validación continua.
- Escalabilidad: La arquitectura propuesta permite crecer hacia soluciones en la nube y análisis en tiempo real.

Modelo sugerido: