Elección de un caso de estudio: Comercio.

Diseño del proyecto

1. **Arquitecturas típicas de proyectos de datos masivos.**

- **Fuentes heterogéneas.**

1. **Instituto Nacional de Estadística e Informática (INEI)**

   Descripción: El INEI proporciona datos estadísticos sobre una variedad de temas, incluido el comercio. Las estadísticas incluyen ventas minoristas, índices de precios al consumidor, y datos sobre importaciones y exportaciones.

   Cómo obtener los datos:

   - **Sitio web:** Visitar el sitio web del INEI.
   - **Descarga directa:** Muchas de las bases de datos están disponibles para descarga directa en formatos como Excel o CSV.

2. **Cámara de Comercio de Lima (CCL)**

   Descripción: La CCL es una organización que agrupa a empresas del sector comercio y proporciona informes, estudios de mercado y datos sobre el desempeño económico de sus miembros.

   Cómo obtener los datos:

   - **Publicaciones:** Revisar las publicaciones periódicas y boletines disponibles en su sitio web oficial.
   - **Afiliación:** Algunos datos pueden estar disponibles solo para miembros, por lo que afiliarse a la CCL podría ser necesario.

3. **Superintendencia Nacional de Aduanas y de Administración Tributaria (SUNAT)**

   Descripción: La SUNAT proporciona datos sobre comercio exterior, importaciones y exportaciones, y el comportamiento de los contribuyentes en el sector comercio.

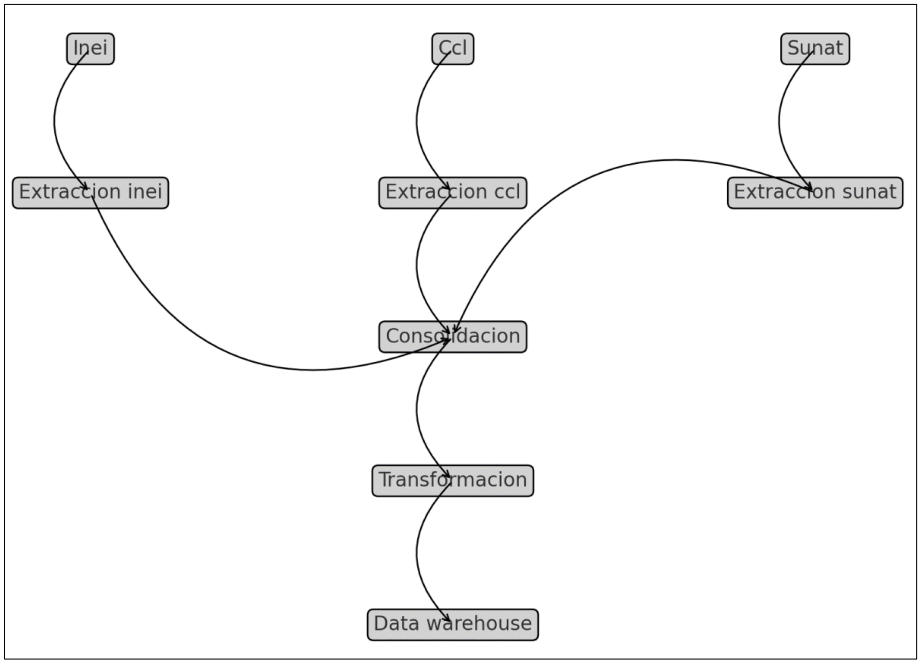
   Cómo obtener los datos:

   - **Portal web:** Visitar el portal de transparencia de SUNAT.

- **Reportes anuales y mensuales:** Descargar reportes estadísticos y boletines mensuales o anuales disponibles en la sección de estadísticas.

► **Extracción, transformación y carga (ETL).**

El proceso ETL se encargará de extraer, transformar y cargar los datos de las tres fuentes identificadas (INEI, CCL y SUNAT) para consolidarlos en un almacén de datos centralizado. A continuación, se presenta el flujo del proceso ETL:



Descripción del Flujo ETL

1. **Extracción:**
   - **INEI:** Descargar los datos estadísticos desde el sitio web del INEI en formato Excel o CSV.
   - **CCL:** Obtener informes y boletines desde el sitio web de la CCL o mediante afiliación.
   - **SUNAT:** Descargar reportes mensuales y anuales desde el portal de transparencia de SUNAT.

2. **Consolidación en Staging:**
   - Crear una zona de staging donde se almacenarán temporalmente los datos extraídos de cada fuente. Esta zona de staging permite realizar preprocesamiento antes de la transformación.

3. **Transformación:**
   - Limpiar y normalizar los datos para asegurar la consistencia entre las diferentes fuentes.

- Realizar transformaciones necesarias, como la conversión de unidades, agregación de datos, y creación de nuevas métricas.
- Integrar los datos en un formato común para su almacenamiento en el Data Warehouse.

4. **Carga:**
   - Cargar los datos transformados en un Data Warehouse centralizado, donde se almacenarán de manera estructurada y organizada para facilitar el análisis y la generación de informes.

- **Herramientas ETL Recomendadas**

1. **Talend:** Es una herramienta ETL de código abierto que permite una fácil integración con múltiples fuentes de datos. Es altamente configurable y tiene una amplia gama de conectores preconstruidos que facilitan la extracción de datos de diferentes fuentes como archivos CSV, Excel, bases de datos, etc.
2. **Apache Nifi: E**s una herramienta ETL que facilita la automatización del flujo de datos entre sistemas. Es ideal para la captura y manipulación de datos en tiempo real y permite la creación de flujos de datos visuales mediante una interfaz de usuario intuitiva.
3. **Microsoft SQL Server Integration Services (SSIS):** Es una poderosa herramienta de integración de datos que forma parte de Microsoft SQL Server. Es adecuada para proyectos de gran escala y permite la extracción, transformación y carga de datos de diversas fuentes de manera eficiente.

- **Almacenamiento**.

Estructuración del Almacenamiento

1. **Esquema:**
   - **Estrella:** Simplifica consultas y optimiza rendimiento.
   - **Copo de Nieve:** Mayor normalización.
2. **Tablas:**
   - **Hechos:** Datos transaccionales (ventas, importaciones).
   - **Dimensiones:** Geografía, productos, clientes, tiempo.
3. **Proceso de Carga:**
   - **Inicial:** Carga masiva de datos históricos.
   - **Actualización:** ETL para datos nuevos y cambios.
   - **Mantenimiento:** Optimización y aseguramiento de integridad.
4. **Optimización:**

- **Índices:** Aceleran consultas.
- **Particionamiento:** Mejora rendimiento.
- **Vistas Materializadas:** Precomputa resultados complejos.

La elección de un Data Warehouse para el almacenamiento de datos transformados es adecuada debido a su capacidad para manejar datos estructurados, su optimización para consultas y análisis complejos, y su capacidad para integrar datos de múltiples fuentes de manera consistente y segura. La estructura del almacenamiento mediante un esquema estrella o copo de nieve, junto con procesos de carga y actualización bien definidos, asegura un rendimiento óptimo y un acceso eficiente a los datos para análisis y toma de decisiones.

- **Tratamiento de los datos.**

  Pasos para Tratar los Datos: Limpieza, Integración y Preparación para el Análisis.

1. **Limpieza de Datos**
   - **Recolección:** Descargar datos de INEI, CCL y SUNAT.
   - **Validación:** Asegurar formatos compatibles y verificar consistencia de datos.
   - **Manejo de Valores Faltantes:** Detectar y tratar valores nulos mediante eliminación, imputación o técnicas avanzadas.
   - **Corrección de Errores:** Identificar y corregir outliers y errores tipográficos.
   - **Normalización y Estandarización:** Escalar variables y uniformar formatos textuales.
2. **Integración de Datos**
   - **Unificación de Fuentes:** Alinear estructuras de datos y mapear campos para integración.
   - **Eliminación de Duplicados:** Identificar y consolidar registros duplicados.
   - **Integración Temporal:** Alinear y agregar datos temporales para coherencia en el análisis.
3. **Preparación para el Análisis**
   - **Transformación:** Crear nuevas variables y calcular agregaciones necesarias.

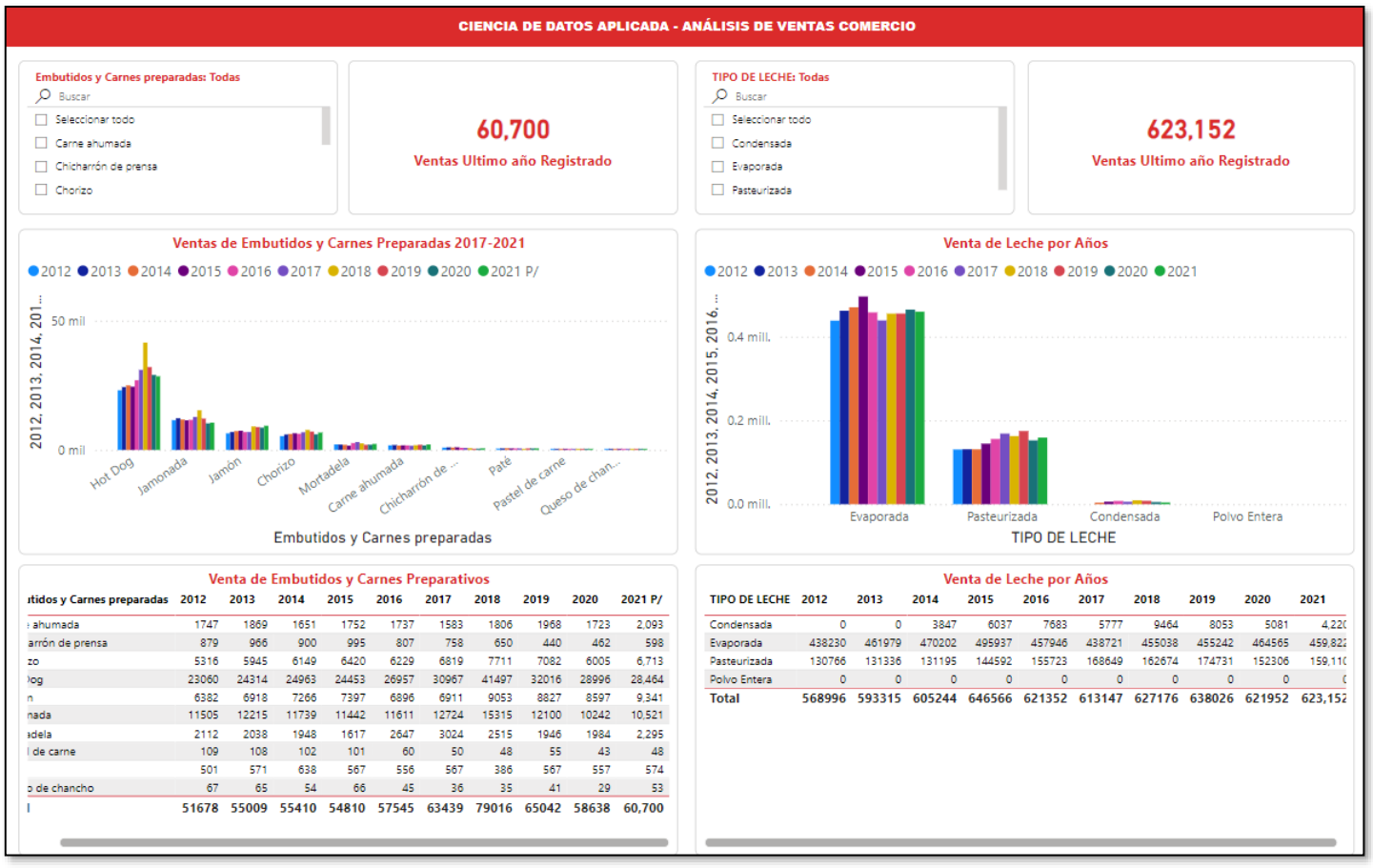
- **Estructuración:** Diseñar un esquema adecuado y cargar datos en el Data Warehouse.
- **Optimización:** Crear índices y particionar tablas grandes para mejorar rendimiento.
- **Validación:** Probar exactitud y consistencia de datos, y verificar integridad en el Data Warehouse.

▸ **Visualización.**

Existen muchas herramientas de visualización de datos, cada una con sus propias ventajas. Algunas son:

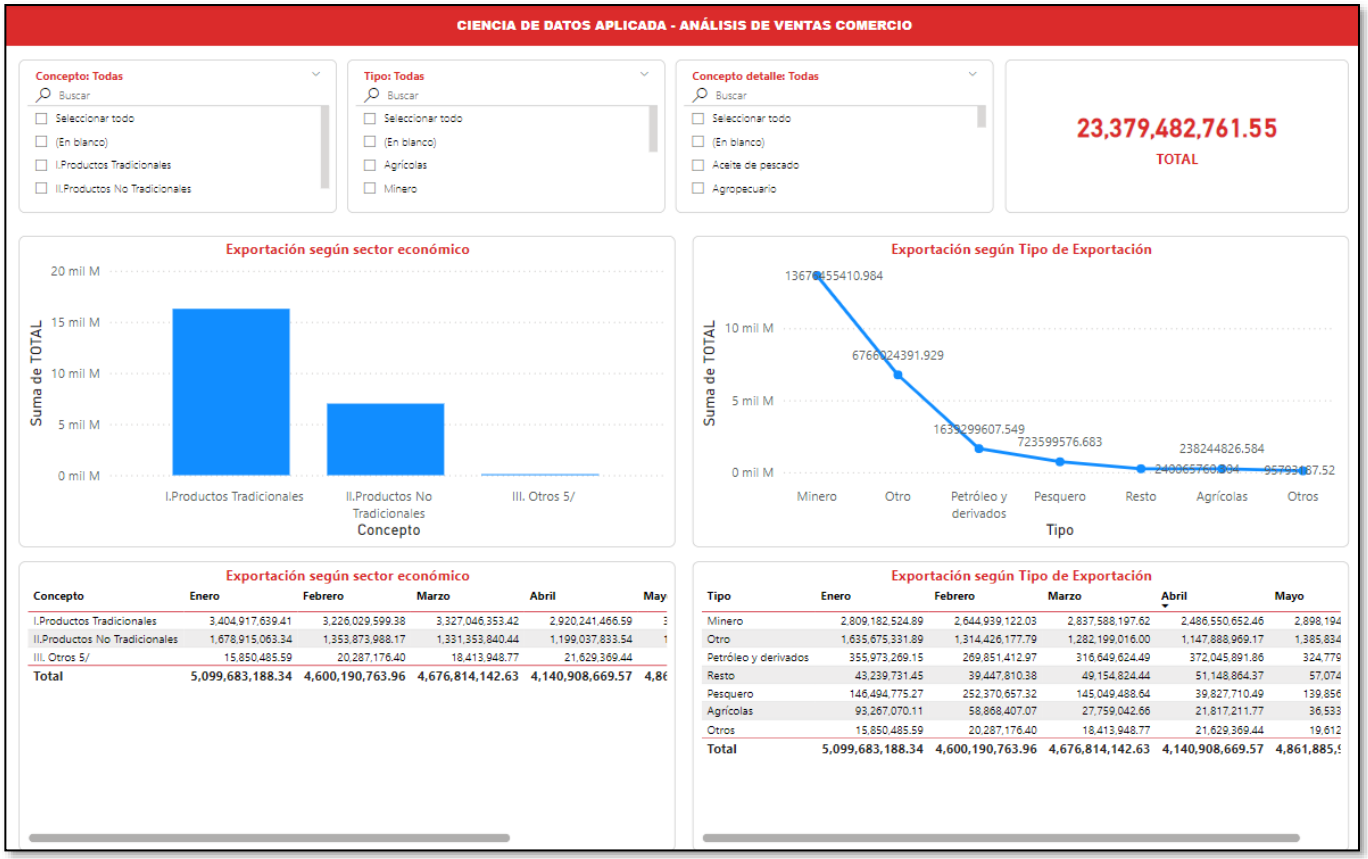
- Tableau: Excelente para crear visualizaciones interactivas y dashboards completos.
- Power BI: Integración fuerte con productos de Microsoft y una buena opción para análisis de datos empresariales.
- Google Data Studio: Herramienta gratuita de Google que permite crear informes y dashboards interactivos.

La herramienta que utilizaremos será Power BI para la visualización de los datos.



tidos y Carnes preparadas	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021 P/
ahumada	1747	1869	1651	1752	1737	1583	1806	1968	1723	2,093
arrón de prensa	879	966	900	995	807	758	650	440	462	598
zo	5316	5945	6149	6420	6229	6819	7711	7082	6005	6,713
log	23060	24314	24963	24453	26957	30967	41497	32016	28996	28,464
n	6382	6918	7266	7397	6896	6911	9053	8827	8597	9,341
nada	11505	12215	11739	11442	11611	12724	15315	12100	10242	10,521
adela	2112	2038	1948	1617	2647	3024	2515	1946	1984	2,295
de carne	109	108	102	101	60	50	48	55	43	48
	501	571	638	567	556	567	386	567	557	574
o de chancho	67	65	54	66	45	36	35	41	29	53
l	51678	55009	55410	54810	57545	63439	79016	65042	58638	60,700

TIPO DE LECHE	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Condensada	0	0	3847	6037	7683	5777	9464	8053	5081	4,22(
Evaporada	438230	461979	470202	495937	457946	438721	455038	455242	464565	459,82
Pasteurizada	130766	131336	131195	144592	155723	168649	162674	174731	152306	159,11(
Polvo Entera	0	0	0	0	0	0	0	0	0	(
Total	568996	593315	605244	646566	621352	613147	627176	638026	621952	623,152



2. **Perfil del científico de datos.**

- **Ciencias de la computación.**

  Habilidades Técnicas Necesarias:

  1. Programación**.**
  2. Manejo de Bases de Datos
  3. Machine Learning**.**
  4. Big Data Technologies
  5. Data Wrangling
  6. Visualización de Datos

  Contribuciones del Equipo:

  - Desarrollador de Software
  - Ingeniero de Datos
  - Científico de Datos.
  - Analista de Datos

- **Matemáticas.**

  1. **Estadística Descriptiva:** Para resumir y describir las características de un conjunto de datos (medias, medianas, modas, desviaciones estándar, etc.).

2. **Inferencia Estadística:** Para hacer predicciones o generalizaciones sobre una población basándose en una muestra.
3. **Regresión Lineal y No Lineal:** Para modelar la relación entre variables dependientes e independientes.
4. **Análisis de Series Temporales:** Para predecir tendencias y patrones en datos temporales.
5. **Métodos de Optimización:** Para encontrar soluciones óptimas en problemas complejos, como el ajuste de hiperparámetros en modelos de machine learning.
6. **Algoritmos de Machine Learning:** Como árboles de decisión, random forest, SVM, k-means, redes neuronales, entre otros.

- **Comunicación.**

1. **Informe Ejecutivo:**
    - **Resumen:** Presentar un resumen ejecutivo claro y conciso de los hallazgos clave y sus implicaciones.
    - **Contexto:** Proveer el contexto del análisis y la metodología utilizada.
    - **Resultados:** Resumir los principales hallazgos y las conclusiones derivadas del análisis.
    - **Recomendaciones:** Proponer acciones específicas basadas en los hallazgos.
2. **Presentaciones Visuales:**
    - **Gráficos y Visualizaciones:** Usar gráficos de barras, líneas, pie charts y otros tipos de visualizaciones para representar datos de manera comprensible.
    - **Infografías:** Crear infografías que resuman los puntos clave de manera visualmente atractiva.
    - **Paneles de Control (Dashboards):** Desarrollar dashboards interactivos que permitan a los usuarios explorar los datos y hallazgos por sí mismos.

- **Negocios.**

1. **Objetivo Estratégico:** Aumentar las ventas y mejorar la experiencia del cliente.
2. **Contribución del Proyecto:** Analizar patrones de compra para optimizar inventarios, diseñar campañas de marketing más efectivas y personalizar la experiencia del cliente.

3. Estrategias en almacenamiento masivo.

- **Data Mart**

Tablas y Datos Incluidos

1. **Tabla de Ventas**
    - **Descripción:** Información detallada sobre cada transacción de venta.
    - **Columnas:**
        - venta_id (Primary Key)
        - fecha_venta
        - cliente_id (Foreign Key)
        - producto_id (Foreign Key)
        - cantidad
        - precio_unitario
        - total_venta
        - metodo_pago

2. **Tabla de Clientes**
    - **Descripción:** Información demográfica y de contacto de los clientes.
    - **Columnas:**
        - cliente_id (Primary Key)
        - nombre
        - apellido
        - genero
        - fecha_nacimiento
        - correo_electronico
        - telefono
        - direccion
        - ciudad
        - pais
        - fecha_registro

3. **Tabla de Productos**
    - **Descripción:** Información detallada sobre los productos vendidos.
    - **Columnas:**
        - producto_id (Primary Key)
        - nombre_producto
        - categoria_id (Foreign Key)
        - precio
        - stock
        - proveedor_id (Foreign Key)

4. **Tabla de Categorías**
    - **Descripción:** Información sobre las categorías de productos.
    - **Columnas:**
        - categoria_id (Primary Key)
        - nombre_categoria
        - descripción

5. **Tabla de Inventarios**
    - **Descripción:** Información sobre los niveles de inventario de productos.
    - **Columnas:**
        - inventario_id (Primary Key)
        - producto_id (Foreign Key)
        - cantidad_disponible
        - fecha_actualizacion

6. **Tabla de Proveedores**
    - **Descripción:** Información sobre los proveedores de productos.
    - **Columnas:**
        - proveedor_id (Primary Key)
        - nombre_proveedor
        - contacto
        - direccion
        - ciudad
        - país

7. **Tabla de Tiempo**
    - **Descripción:** Dimensión temporal para análisis de ventas a lo largo del tiempo.
    - **Columnas:**
        - fecha (Primary Key)
        - dia
        - mes
        - año
        - trimestre
        - dia_semana
        - es_fin_de_semana

Relaciones entre las Tablas

- Ventas está relacionada con Clientes a través de cliente_id.
- Ventas está relacionada con Productos a través de producto_id.
- Productos está relacionada con Categorías a través de categoria_id.
- Productos está relacionada con Proveedores a través de proveedor_id.
- Inventarios está relacionada con Productos a través de producto_id.
- Ventas está relacionada con Tiempo a través de fecha_venta.

- **Data Warehouse.**

Componentes del Data Warehouse

1. **ETL (Extract, Transform, Load)**
    - **Extract:** Extracción de datos desde múltiples fuentes.
    - **Transform:** Limpieza y transformación de datos.

- **Load:** Carga de datos transformados en el Data Warehouse.

2. **Área de Staging**
   - **Descripción:** Almacenamiento temporal de datos extraídos antes de la transformación y carga final.
   - **Componentes:** Tablas intermedias con datos sin procesar.
3. **Data Warehouse Central**
   - **Descripción:** Almacenamiento principal de datos transformados y normalizados.
   - **Componentes:** Tablas de hechos y dimensiones organizadas.
4. **Data Marts**
   - **Descripción:** Subconjuntos del Data Warehouse para áreas específicas del negocio.
   - **Componentes:** Tablas optimizadas para análisis específicos.

- **Integración con el Data Mart**

1. **Estructura del Data Warehouse**
   - **Tablas de Hechos:**
     - **Hechos de Ventas:** fact_ventas (venta_id, fecha_id, cliente_id, producto_id, tienda_id, cantidad, precio_unitario, total_venta, metodo_pago)
     - **Hechos de Inventarios:** fact_inventarios (inventario_id, producto_id, fecha_id, cantidad_disponible)
   - **Tablas de Dimensiones:**
     - **Dimensión de Tiempo:** dim_tiempo (fecha_id, fecha, dia, mes, año, trimestre, dia_semana, es_fin_de_semana)
     - **Dimensión de Clientes:** dim_clientes (cliente_id, nombre, apellido, genero, fecha_nacimiento, correo_electronico, telefono, direccion, ciudad, pais, fecha_registro)
     - **Dimensión de Productos:** dim_productos (producto_id, nombre_producto, categoria_id, precio, proveedor_id)
     - **Dimensión de Categorías:** dim_categorias (categoria_id, nombre_categoria, descripcion)
     - **Dimensión de Proveedores:** dim_proveedores (proveedor_id, nombre_proveedor, contacto, direccion, ciudad, pais)
     - **Dimensión de Tiendas:** dim_tiendas (tienda_id, nombre_tienda, ubicacion, gerente).

2. **Proceso de ETL**
    - **Extracción:** Recopilación de datos de fuentes internas y externas.
    - **Transformación:** Limpieza y normalización de datos.
    - **Carga:** Inserción de datos en las tablas del Data Warehouse.
3. **Integración con el Data Mart**
    - **Selección de Datos Relevantes:** Filtrado y transformación de datos para llenar Data Marts específicos.
    - **Optimización para Análisis Específicos:** Estructuración de Data Marts para consultas rápidas y análisis.
    - **Actualización Regular:** Actualización periódica de Data Marts para reflejar datos recientes del Data Warehouse.

- **Data Lake.**

1. Almacenamiento Escalable y Económico

- **Escalabilidad:** Los Data Lakes permiten el almacenamiento de grandes volúmenes de datos de diferentes tipos sin necesidad de estructura previa. Esto es ideal para el sector comercio, que genera datos masivos y diversos (transacciones, datos de clientes, inventarios, etc.).

2. Flexibilidad en el Tipo de Datos

- **Datos Estructurados y No Estructurados:** Pueden almacenar datos estructurados (bases de datos relacionales), semi-estructurados (JSON, XML) y no estructurados (imágenes, videos, logs).

3. Acceso Rápido y Procesamiento Paralelo

- **Acceso Directo:** Los analistas y científicos de datos pueden acceder directamente a los datos en bruto para realizar análisis exploratorios y desarrollar modelos de machine learning.

4. Facilita el Data Discovery y la Innovación

- **Exploración de Datos:** Facilita la exploración y el descubrimiento de nuevas tendencias y patrones en los datos sin la necesidad de preprocesarlos.

5. Integración con Herramientas Analíticas y de Machine Learning

- **Compatibilidad:** Se integra fácilmente con herramientas analíticas y plataformas de machine learning como AWS Sagemaker, Google AI Platform, Databricks, etc.

Gestión de la Ingesta de Datos No Estructurados

1. Proceso de Ingesta de Datos

- **Fuentes de Datos:** Identificación de todas las fuentes de datos no estructurados como archivos de registro (logs), redes sociales, correos electrónicos, imágenes, videos y documentos.
- **ETL/ELT Pipeline:**
  - **Extract:** Recolección de datos no estructurados desde diferentes fuentes utilizando conectores específicos (API, FTP, web scraping, etc.).
  - **Transform:** Aplicación de procesos de limpieza y preprocesamiento según sea necesario (eliminación de duplicados, corrección de errores).
  - **Load:** Almacenamiento de los datos en su formato original en el Data Lake, usando tecnologías como AWS S3, Azure Blob Storage, Google Cloud Storage, etc.

2. Organización y Catalogación de Datos

- **Data Catalog:** Uso de un catálogo de datos para indexar y catalogar todos los datos ingresados, facilitando la búsqueda y acceso a los datos.
- **Metadata Management:** Gestión de metadatos para describir la estructura, origen y características de los datos no estructurados.

3. Preprocesamiento y Enriquecimiento de Datos

- **Data Wrangling:** Uso de herramientas de data wrangling para limpiar y transformar los datos no estructurados en formas utilizables para análisis posteriores.
- **Enriquecimiento:** Integración de datos no estructurados con datos estructurados para obtener una visión más completa del negocio.

4. Seguridad y Gobernanza de Datos

- **Políticas de Acceso:** Implementación de políticas de acceso y control para asegurar que solo los usuarios autorizados puedan acceder a datos sensibles.
- **Compliance:** Aseguramiento del cumplimiento de normativas y regulaciones de privacidad y seguridad de datos (GDPR, CCPA, etc.).

► **Nuevas tendencias en almacenamiento masivo.**

Almacenamiento Descentralizado en Blockchain

El almacenamiento descentralizado en blockchain utiliza una red distribuida de nodos para almacenar datos en lugar de depender de un solo proveedor centralizado. Los datos se dividen en fragmentos encriptados y se distribuyen a través de múltiples

nodos en la red. Esta tecnología ofrece varias ventajas sobre los métodos tradicionales y el almacenamiento en la nube centralizado.

Ventajas:

- Seguridad y Privacidad: La encriptación de los datos garantiza que solo el propietario de los datos tenga acceso a ellos.
- Redundancia y disponibilidad: Al estar distribuidos en múltiples nodos, los datos tienen redundancia incorporada, lo que mejora la disponibilidad y la durabilidad.
- Costos reducidos: Los costos pueden ser más bajos ya que no se necesita una infraestructura centralizada costosa.
- Ideal para el almacenamiento de grandes cantidades de datos que necesitan ser accesibles durante largos períodos de tiempo sin riesgo de pérdida o manipulación.
- Los registros que requieren alta seguridad y privacidad pueden beneficiarse enormemente de la inmutabilidad y seguridad de la blockchain.

El almacenamiento descentralizado en blockchain representa una evolución significativa en la forma en que gestionamos y almacenamos grandes cantidades de datos, ofreciendo una alternativa segura, escalable y económica a las soluciones tradicionales y en la nube centralizadas.

4. Estrategias de aplicación de la ciencia de datos y datos masivos.

- **Inteligencia de negocio.**
  - **Recolección y Consolidación de Datos**
    - ✓ **Fuente de Datos**: Información de SUNAT, INEI y CCL.
    - ✓ **Almacenamiento**: Data Warehouse (Amazon Redshift) para datos limpios y transformados.
    - ✓ **Beneficio**: Visión integral de las operaciones comerciales y económicas.
  - **Análisis Descriptivo**
    - ✓ **Herramientas**: Tableau, Power BI.
    - ✓ **Visualizaciones**: Dashboards interactivos que muestran ventas, tendencias de importación/exportación, y desempeño de productos.
    - ✓ **Beneficio**: Visualización rápida y eficiente de datos clave para identificar patrones y tendencias.
  - **Análisis Predictivo**
    - ✓ **Técnicas**: Regresión lineal, modelos de series temporales.

- ✓ **Aplicación**: Predicción de demanda de productos, pronósticos de ventas.
- ✓ **Beneficio**: Anticipar cambios en el mercado y ajustar estrategias comerciales proactivamente.

- **Análisis Prescriptivo**
  - ✓ **Herramientas**: Algoritmos de optimización y simulación.
  - ✓ **Aplicación**: Optimización de inventarios, estrategias de precios dinámicos.
  - ✓ **Beneficio**: Recomendaciones concretas para mejorar eficiencia operativa y maximizar beneficios.
- **Minería de Datos**
  - ✓ **Técnicas**: Clustering, análisis de asociación.
  - ✓ **Aplicación**: Identificación de segmentos de mercado, análisis de cesta de la compra.
  - ✓ **Beneficio**: Descubrimiento de relaciones ocultas para estrategias de marketing y ventas.
- **Monitoreo Continuo**
  - ✓ **Implementación**: Sistemas de monitoreo en tiempo real.
  - ✓ **Aplicación**: Alertas sobre desviaciones significativas en ventas, cambios abruptos en tendencias de mercado.
  - ✓ **Beneficio**: Respuestas rápidas a eventos inesperados, minimizando riesgos y aprovechando oportunidades emergentes.
- **Mejora de Procesos Internos**
  - ✓ **Análisis de Desempeño**: Evaluación de eficiencia en la cadena de suministro, tiempos de entrega.
  - ✓ **Beneficio**: Eliminación de cuellos de botella, mejorando eficiencia y reduciendo costos operativos.
- **Estrategia de Personalización**
  - ✓ **IA y BI**: Modelos de inteligencia artificial para analizar el comportamiento de los clientes.
  - ✓ **Aplicación**: Campañas de marketing personalizadas basadas en preferencias y comportamientos de compra.
  - ✓ **Beneficio**: Aumenta la satisfacción del cliente y las tasas de conversión.
- **Implementación Práctica**
  - ✓ **Recopilación de Datos**: Datos extraídos de SUNAT, INEI y CCL e integrados en el Data Warehouse.
  - ✓ **Análisis en Tiempo Real**: Uso de herramientas de BI para análisis y generación de reportes y dashboards.

- ✓ **Modelos Predictivos y Prescriptivos**: Desarrollo de modelos que proporcionan predicciones y recomendaciones.
- ✓ **Feedback Loop**: Implementación de recomendaciones, monitoreo de impacto y ajuste continuo de modelos y estrategias.

▸ **Analítica de negocio.**

1. Definición del Problema y Objetivos

Problema a abordar:

- **Identificar:** Oportunidades de crecimiento y problemas operativos en el comercio de productos de consumo masivo.

Objetivos del análisis:

- **Oportunidades:** Identificar productos con alta demanda y mercados emergentes.
- **Problemas:** Detectar ineficiencias en la cadena de suministro y caídas en ventas.

2. Recopilación y Preparación de Datos

Fuentes de datos:

- **SUNAT:** Datos de importaciones y exportaciones.
- **INEI:** Estadísticas de comercio y censos económicos.
- **CCL:** Informes del sector comercio y datos de afiliados.

Preparación de datos:

- **Limpieza:** Eliminación de duplicados, manejo de valores nulos.
- **Integración:** Unificación de formatos y criterios de codificación.
- **Enriquecimiento:** Agregar variables calculadas (e.g., tasa de crecimiento mensual).

3. Análisis Descriptivo

Análisis inicial:

- **Desempeño de ventas:**
  - ✓ **Métricas:** Total de ventas mensuales, promedio de ventas diarias.
  - ✓ **Herramientas:** Gráficos de línea y de barras en Tableau o Power BI.
- **Análisis de productos:**
  - ✓ **Métricas:** Productos más vendidos, margen de beneficio por producto.
  - ✓ **Herramientas:** Gráficos de Pareto, análisis ABC.
- **Visualizaciones:**
  - ✓ **Dashboard interactivo:** Ventas por región, productos con mayor crecimiento, tendencias de importación/exportación.

4. Análisis Predictivo

Modelos predictivos:

- **Predicción de demanda:**
  - ✓ **Modelo:** Series temporales (ARIMA).
  - ✓ **Datos:** Históricos de ventas.
- **Proyección de ventas:**
  - ✓ **Modelo:** Regresión lineal múltiple.
  - ✓ **Datos:** Ventas históricas, datos económicos (e.g., PIB, inflación).

5. Análisis Prescriptivo

Recomendaciones de acciones:

- **Optimización de inventarios:**
  - ✓ **Modelo:** Algoritmo de optimización (e.g., programación lineal).
  - ✓ **Aplicación:** Determinar niveles óptimos de inventario para productos clave.
- **Estrategias de precios dinámicos:**
  - ✓ **Modelo:** Análisis de elasticidad de precios.
  - ✓ **Aplicación:** Ajuste de precios basado en demanda y competencia.

6. Identificación de Oportunidades

Segmentación de mercado:

- ✓ **Clusterización:** Uso de K-means para segmentar clientes y mercados.
- ✓ **Identificación:** Mercados emergentes y segmentos de alto valor.

Análisis de productos:

- ✓ **Análisis de asociación:** Reglas de asociación (e.g., algoritmo Apriori).
- ✓ **Aplicación:** Identificación de productos complementarios para ventas cruzadas.

7. Identificación de Problemas

Análisis de cadena de suministro:

- ✓ **Métricas:** Tiempo de entrega, costos logísticos.
- ✓ **Herramientas:** Análisis de procesos, simulación de escenarios.
- ✓ **Problema:** Identificación de cuellos de botella y oportunidades de mejora.

Análisis de ventas:

- ✓ **Tendencias negativas:** Detección de productos con ventas decrecientes.
- ✓ **Causas:** Análisis de factores externos (e.g., cambios en la demanda, competencia).

8. Comunicación de Resultados

Informe ejecutivo:

- ✓ **Contenido:** Resumen de hallazgos, visualizaciones clave, recomendaciones.
- ✓ **Formato:** PDF, presentación en PowerPoint.

Presentación visual:

- ✓ **Audiencia:** Ejecutivos y tomadores de decisiones.
- ✓ **Formato:** Dashboards interactivos, gráficos claros y concisos.

Ejemplo de Dashboard

Secciones del dashboard:

- **Ventas totales:**
  - ✓ **Gráfico de líneas:** Tendencia de ventas mensuales.
- **Productos más vendidos:**
  - ✓ **Gráfico de barras:** Top 10 productos por ventas.
- **Desempeño regional:**
  - ✓ **Mapa de calor:** Ventas por región.
- **Análisis de cadena de suministro:**
  - ✓ **Gráfico de Gantt:** Tiempos de entrega y puntos críticos.
- **Predicción de demanda:**
  - ✓ **Gráfico de líneas:** Proyección de ventas futuras.

- **Minería de datos.**

Pasos para la Minería de Datos

1. **Recolección de Datos**
   - **Fuentes de Datos**: Bases de datos internas de ventas, inventarios, encuestas de clientes, sistemas de punto de venta, etc.
   - **Formatos de Datos**: CSV, Excel, SQL, etc.

2. **Preparación de Datos**
   - **Limpieza de Datos**: Eliminar duplicados, manejar valores faltantes, corregir errores.
   - **Transformación de Datos**: Normalización, creación de nuevas variables, codificación de variables categóricas.
   - **Integración de Datos**: Combinar datos de diferentes fuentes para obtener una vista consolidada.

3. **Exploración de Datos**
   - **Análisis Descriptivo**: Resúmenes estadísticos, visualización de datos (gráficos de barras, histogramas, diagramas de caja, etc.).

- **Detección de Patrones Iniciales**: Identificación de correlaciones y relaciones entre variables.

4. **Modelado**
   - **Técnicas de Minería de Datos**:
     - ✓ **Análisis de Clúster**: Para segmentar clientes o productos en grupos similares.
     - ✓ **Regresión**: Para identificar relaciones y predecir ventas futuras.
     - ✓ **Árboles de Decisión**: Para clasificar y predecir comportamientos.
     - ✓ **Análisis de Asociación**: Para descubrir relaciones entre productos (por ejemplo, análisis de cesta de la compra).
     - ✓ **Series Temporales**: Para analizar tendencias a lo largo del tiempo.
5. **Evaluación**
   - **Validación del Modelo**: Utilizar técnicas como validación cruzada para evaluar la precisión del modelo.
   - **Interpretación de Resultados**: Comprender y comunicar los hallazgos.
6. **Implementación**
   - **Despliegue del Modelo**: Integración del modelo en sistemas operativos para toma de decisiones en tiempo real.
   - **Monitorización y Mantenimiento**: Asegurar que el modelo sigue siendo preciso y relevante.

Herramientas Comunes

- **Python** (bibliotecas como Pandas, Scikit-learn, Matplotlib, Seaborn)
- **R** (paquetes como dplyr, ggplot2, caret)
- **Software de Minería de Datos**: RapidMiner, KNIME
- **Bases de Datos**: SQL, NoSQL
- **Plataformas de BI**: Power BI, Tableau

- **Aprendizaje automático.** Se utilizó la base de datos la baje de kaggle de las ventas diarias del queso cheddar

```
!pip install openpyxl
# Importar panda
import pandas as
# Leer el Archivo Excel
df = pd.read_excel('/content/Cheddar Cheese Prices and Sales.xlsx', engine='openpyxl')
```

```
Requirement already satisfied: openpyxl in /usr/local/lib/python3.10/dist-packages (3.1.5)
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.10/dist-packages (from openpyxl) (1
```

```
# Preprocesamiento Inicial
# Verifica las primeras filas para entender mejor tus datos
print(df.head())
```

```
  Week Ending Date Report Date   Date  Weighted Prices     Sales
0       08/19/2017  08/23/2017  07/22           1.5907  11217751
1       08/19/2017  08/23/2017  07/29           1.6226  11933852
2       08/19/2017  08/23/2017  08/05           1.6822  11874522
3       08/19/2017  08/23/2017  08/12           1.7376  11228718
4       08/19/2017  08/23/2017  08/19           1.7429  12414048
```

```
# Obtén información sobre tipos de datos y valores faltantes
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1410 entries, 0 to 1409
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Week Ending Date  1410 non-null   object
 1   Report Date       1410 non-null   object
 2   Date              1410 non-null   object
 3   Weighted Prices   1410 non-null   float64
 4   Sales             1410 non-null   int64
dtypes: float64(1), int64(1), object(3)
memory usage: 55.2+ KB
None
```

```
# Conviertir 'Week Ending Date' a datetime para usarlo como referencia temporal
df['Week Ending Date'] = pd.to_datetime(df['Week Ending Date'])
```

```
# Nos Aseguramos de que no haya valores faltantes
df.dropna(inplace=True)
# Verifica que no haya valores nulos o duplicados (opcional dependiendo del dataset)
print(df.isnull().sum())
df = df.drop_duplicates()
```

```
Week Ending Date    0
Report Date         0
Date                0
Weighted Prices     0
Sales               0
dtype: int64
```

```
# Agrupar las ventas por mes sumando los valores diarios para obtener totales mensuales
df_monthly_sales = df.set_index('Week Ending Date').resample('M')['Sales'].sum().reset_index()
```

```
# Dividir los Datos en Conjuntos de Entrenamiento y Prueba, 3 meses
N = 3 # Ejemplo: Reservar los últimos tres meses para pruebas

train_data = df_monthly_sales[:-N]
test_data = df_monthly_sales[-N:]

X_train = train_data[['Week Ending Date']]
y_train = train_data['Sales']

X_test = test_data[['Week Ending Date']]
y_test = test_data['Sales']
```

```
# Transformar las Fechas en Características Numéricas

X_train['Year'] = X_train['Week Ending Date'].dt.year
X_train['Month'] = X_train['Week Ending Date'].dt.month

X_test['Year'] = X_test['Week Ending Date'].dt.year
X_test['Month'] = X_test['Week Ending Date'].dt.month

# Descartar la columna 'Week Ending Date'
X_train.drop(['Week Ending Date'], axis=1, inplace=True)
X_test.drop(['Week Ending Date'], axis=1, inplace=True)


# Crear y Entrenar un Modelo Predictivo

from sklearn.linear_model import LinearRegression

model = LinearRegression()
model.fit(X_train, y_train) # Asegúrate de entrenarlo con características numéricas apropiadas.
```

```
▾ LinearRegression
LinearRegression()
```

```
# Si 'Year' y 'Month' son las únicas características después del preprocesamiento:
model.fit(X_train[['Year', 'Month']], y_train)
```

```
▾ LinearRegression
LinearRegression()
```

- **Inteligencia artificial.**

1. **Análisis Predictivo Avanzado**

   Aplicaciones:

   - Predicción de Ventas: Utilizar redes neuronales profundas para prever ventas futuras considerando múltiples variables.
   - Predicción de Demanda: Anticipar demanda de productos por región y temporada para optimizar inventarios y reducir costos.

   Beneficios:

   - Mejora en la planificación estratégica y operativa.
   - Reducción de costos y aumento de ingresos.

2. **Análisis de Sentimiento y Retroalimentación del Cliente**

   Aplicaciones:

   - Análisis de Opiniones: Usar NLP para evaluar comentarios y reseñas de clientes en tiempo real.
   - Chatbots y Asistentes Virtuales: Interactuar con clientes para recopilar opiniones y ofrecer soporte.

Beneficios:

- Mejora de la experiencia del cliente.
- Identificación rápida de problemas y oportunidades.

3. **Optimización de Precios**

   Aplicaciones:

   - Modelos de Precios Dinámicos: Ajustar precios en tiempo real basados en demanda y competencia.
   - Descuentos Personalizados: Ofrecer descuentos basados en el comportamiento de compra del cliente.

   Beneficios:

   - Maximización de ingresos.
   - Aumento de la competitividad.