**Trabajo Práctico**

La compañía "CDD-IN" nació como un almacén de barrio, y a través de los años se convirtió en una importante cadena de ventas de productos a nivel nacional e internacional.

Este crecimiento se vio acompañado con la adquisición de software para el manejo de diversos procesos de negocio, como ser: Ventas, Stock, RRHH. La mayor parte de los datos están alojados en una base de datos relacional. No obstante, aún existen algunas sucursales que aportan sus datos en planillas Excel y archivos TXT.

Por último, existen algunas sucursales muy pequeñas que aportan sus informes utilizando el correo electrónico

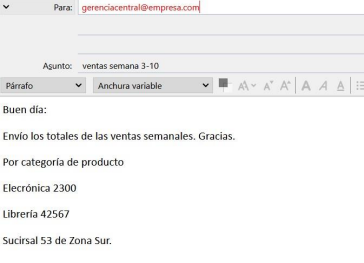
**Problemática**

Al día de la fecha, están buscando mejorar su proceso de ventas, motivo por el cual requiere aumentar la cantidad de reportes del área y la frecuencia con la cual se generan. Hoy en día, dichos reportes se manejan en archivos Excel, que implican un **excesivo costo operativo** en la construcción y **alta probabilidad de errores**. A esto se le suma el problema de la **baja capacidad de respuesta** ante cambios solicitados por los usuarios finales.

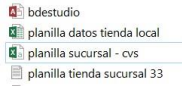
Los usuarios finales son del nivel estratégico, con tendencia a realizar análisis macro de diversas aristas de la información, realizando comparaciones para evaluar la evolución de sus métricas. Entre ellas ventaspor Producto, empleado, sucursal y mes, así como también los costos por Empleado, Región y sucursal.

Por otro lado, existen algunos de ellos que suelen profundizar en sus análisis, y que debido a que disponen de cierto expertise técnico, toman los datos que se les envían y realizan sus propias gráficas en busca de descubrir conocimiento, aunque con frecuencia se quedan del poco tiempo disponible y las limitaciones de las herramientas con las que cuentan. Entre ellas la más frecuente es la venta de productos asociados en el mismo ticket o factura de ventas. También, se desea hacer un análisis para encontrar cual es el producto más vendido por categorías y segmentación de edades según los siguientes tramos: 18 a 30, 31 a 45, 46 a 60 y 60 a 75 años.

Para ambos tipos de análisis, si bien se cuenta con una base de datos que dispone de datos gobernados, existen algunos datos que no están en un sistema transaccional, por lo cual se manejan en archivos txt, xls o correo del tipo:



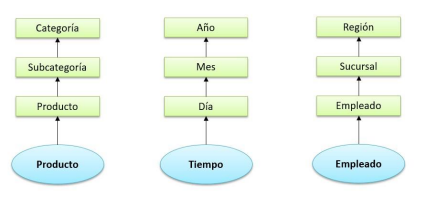
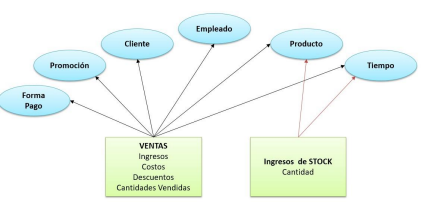
Algunos archivos son

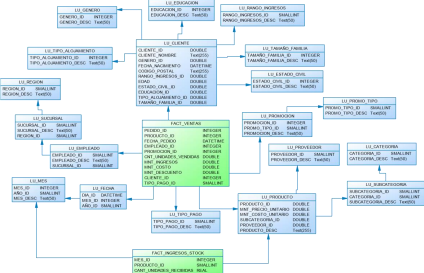


**Información disponible**

**Modelo de datos**

La empresa dispone de una base de datos, que si bien no está finalizado, contiene información de las ventas y stock Los datos disponibles son los siguientes:





Tareas 1: Fuentes

Analizar las fuentes y datos y clasificarlos utilizando el siguiente cuadro.

|  |  |  |  |
| --- | --- | --- | --- |
| Nombre | Fuente | Tipo | Descripción |
| BDEmpresa | Interna | Base de datos relacional Access | Es una fuente de datos de tipo interna ya que en ella se almacenan gran parte de los datos sobre las ventas de la sucursal. |
| Planilla de Excel | Interna | Planilla de datos de sucursal- Estructurado | Es una fuente de datos en la cual se guardan los datos sobre las ventas.  Y está conformado por distintos datos separados por punto y coma |
| Archivos TXT | Interna | Archivos de informes con datos no estructurados | Es una fuente de datos interna debido a que cada sucursal informa los datos de venta correspondientes.  Usando como separador de los datos el punto y coma. |

Conclusión acerca de las fuentes de datos

Se pudo llegar a la conclusión de que las distintas fuentes de datos son del tipo interna ya que almacenan los registros de información sobre los datos de los distintos sectores de la empresa. Además hay planillas que son también del tipo estructuradas que almacenan datos sobre las distintas ventas. Por otro lado hay una fuente con datos del tipo no estructuradas porque son archivos de texto que se utilizan para enviar ciertos informes de ventas pero que no respetan un formato en específico.

*Tarea 2: Tipos de datos*

De los datos expresados anteriormente en el modelo, ¿podría especificar a qué tipo corresponden? Es decir,

• ¿Cuáles de ellos son métricas y cuáles dimensiones o aperturas? • ¿Cuáles son las tablas de hechos?

• ¿Qué jerarquías identifica?

• En la unión entre FACT y LOOKUP, ¿encuentra algún inconveniente en los campos?

• ¿Existe algún campo evitable en la tabla LU\_CLIENTE?

**Desarrollo**

**BD Empresa:** Está constituido en un tipo de dato estructurado ya que está desarrollado con un modelo relacional.

**Planilla Excel:** Está constituido en un tipo de dato estructurado ya que la grilla tiene un formato específico.

**Archivos TXT:** Está constituido en un tipo de dato no estructurado. Ya que hay algunos datos o columnas que no se cargan.

\*Métricas:

-Costo

-Venta

\*Dimensiones

-Sucursal

-Método de pago

-Categoría

-tipo de promoción

Tablas de hechos:

\*fact\_ventas

Las jerarquías que se identificaron fueron la de fecha de pedido y la de las sucursales que pueden estar asociadas:

\*Fecha Venta => año => mes => día.

\*Sucursal => región.

El inconveniente que se encontró entras las tablas FACT y lookup fue que la columna PEDIDO\_ID no tiene ninguna relación con otra tabla. Si no que es una columna para la realización de un detalle.

En la tabla LU\_CLIENTE se podrían obviar los siguientes campos.  
\*EDUCACION\_ID: porque sería un dato redundante ya que no haría falta saber el nivel académico del cliente.

\*TIPO\_ALOJAMIENTO: porque no haría falta saber en donde se aloja el cliente

\*ESTADO\_CIVIL y TAMAÑO\_FAMILIA: porque no es necesario saber si el cliente está casado y/o divorciado o si posee hijos.

\*GENERO: Ya que no es sumamente necesario saber el sexo del cliente para realizar una venta.

\*CLIENTE\_ID: Ya que es un dato que sirve simplemente para la base de datos

\*EDAD: ya que se lo puede calcular a través del campo de FECHA\_NAC.  
\*CODIGO\_POSTAL: ya que no se podría transformar los datos para obtener la descripción.

Los campos válidos serían:

\*FECHA\_NAC

\*RANGO\_INGRESOS

\*CLIENTE\_NOMBRE

*Tarea 3: Objetivos*

Establezca los objetivos del trabajo, es decir las líneas de trabajo hacia las que deben conducir los esfuerzos.

**Objetivos del Trabajo**

* Migrar todos los datos a una base de datos.
* Mejorar Procesos de ventas.
  + Mayor cantidad de reportes y con mayor frecuencia
  + Automatizar el proceso de creación de los reportes.

*Tarea 4: Procesos ETL*

Realizar el análisis de cada fuente de datos y decidir qué tipo de proceso se va a utilizar (ETL, ELT, etc), Justificar

¿Qué acciones se realizaron sobre datos irrelevantes, duplicados, no estructurados y ambiguos?

Diseñar el modelo de datos y los procesos (algoritmos) de extracción, transformación si fuera necesario, y carga en el DW.

**Desarrollo:**

**Para poder trabajar con los datos utilizamos el proceso ETL (extract,transform, load), esto se debe al procedimiento de carga de datos, como son diferentes tipos de archivos los cuales contienen los datos de la empresa, se procedió a extraer dichos datos, transformarlos convirtiéndolos todos a formato de access con el propósito de poder trabajar con ellos en un solo formato y por medio de una única aplicación y se procedió a exportar dichos datos.**

**En este caso, procedimos a convertir o exportar todos los archivos de datos en Access (de office), con excepción del archivo TXT ya que, contenía la misma información que otro archivo, por ende, podemos “despreciar” ese archivo de datos.**

**Luego utilizando SQL para realizar las consultas que nos permiten analizar la información que tenemos a disposición.**

*Tarea 5: Data Warehouse*

Se requiere el diseño del Data Warehouse para analizar las cantidades vendidas y el ingreso por ventas por Producto, empleado, sucursal y mes, así como también los costos por Empleado, Región y sucursal. También se deberían contemplar los otros dos requerimientos: Segmentación y asociación de productos.

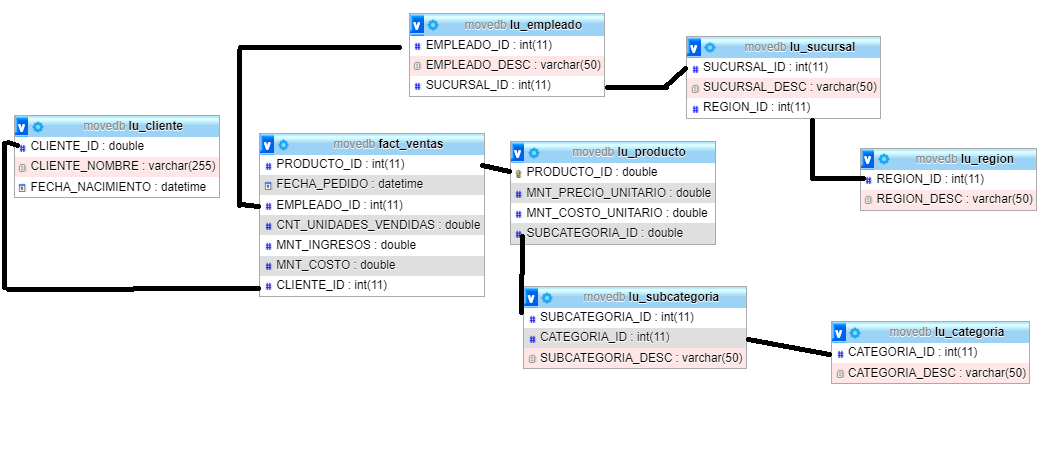
¿Qué modelo utilizaría? ¿Copo de nieve o estrella? ¿Trabajará con varios Data marts o con uno solo? Justifique ambas elecciones y representa el modelo.

Desarrollo

El modelo utilizado es del tipo copo de nieve. Elegimos este modelo porque es el que mejor se adapta a las fuentes de datos que ya teníamos y nos permite obtener los datos que necesitamos sin hacer muchos cambios en la base de datos.

Utilizamos un solo data mart ya que los datos que necesitamos están en una sola base de datos.

La herramienta que usamos para migrar los datos de access a mysql se llama Bullzip MS Access to MySQL.

**

Ventas por producto

[SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) p1.PRODUCTO\_DESC, ROUND([sum](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/group-by-functions.html" \l "function_sum)(v1.MNT\_COSTO), 2) FROM lu\_producto p1, fact\_ventas v1 WHERE p1.PRODUCTO\_ID = v1.PRODUCTO\_ID GROUP BY v1.PRODUCTO\_ID;

ventas por empleado

[SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) e1.EMPLEADO\_DESC, ROUND([sum](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/group-by-functions.html" \l "function_sum)(v1.MNT\_COSTO), 2) FROM fact\_ventas v1, lu\_empleado e1 WHERE v1.EMPLEADO\_ID = e1.EMPLEADO\_ID GROUP BY e1.EMPLEADO\_ID;

ventas por sucursal y mes

[SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) s1.SUCURSAL\_DESC, ROUND([sum](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/group-by-functions.html" \l "function_sum)(v1.MNT\_COSTO), 2) FROM fact\_ventas v1, lu\_empleado e1, lu\_sucursal s1 WHERE v1.EMPLEADO\_ID = e1.EMPLEADO\_ID [AND](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/logical-operators.html#operator_and) e1.SUCURSAL\_ID = s1.SUCURSAL\_ID GROUP BY s1.SUCURSAL\_ID;

ventas por mes

SET lc\_time\_names = 'es\_ES';

[SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) [sum](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/group-by-functions.html#function_sum)(v1.MNT\_INGRESOS), MONTHNAME(v1.FECHA\_PEDIDO),[YEAR](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/date-and-time-types.html)(v1.FECHA\_PEDIDO) FROM fact\_ventas v1 GROUP BY v1.FECHA\_PEDIDO;

costos por empleado región y sucursal

no se puede

Venta de productos asociados en el mismo ticket o factura de ventas

[SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) p1.PRODUCTO\_DESC , v1.PEDIDO\_ID FROM fact\_ventas v1, lu\_producto p1 WHERE v1.PRODUCTO\_ID = p1.PRODUCTO\_ID ORDER BY v1.PEDIDO\_ID;

Producto más vendido por categorías y segmentaciones de edades según los siguientes tramos de edades:

[SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) p1.PRODUCTO\_DESC , c1.CATEGORIA\_DESC FROM fact\_ventas v1, lu\_producto p1, lu\_subcategoria s1, lu\_categoria c1 WHERE v1.PRODUCTO\_ID = p1.PRODUCTO\_ID [AND](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/logical-operators.html#operator_and) p1.SUBCATEGORIA\_ID = s1.SUBCATEGORIA\_ID [AND](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/logical-operators.html#operator_and) s1.CATEGORIA\_ID = c1.CATEGORIA\_ID GROUP BY c1.CATEGORIA\_DESC;

\*18 a 30

[SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) sub01.producto, sub01.edad, sub01.cant FROM ( [SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) [COUNT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/group-by-functions.html#function_count)(p1.PRODUCTO\_ID) as cant, p1.PRODUCTO\_DESC AS producto, TIMESTAMPDIFF([YEAR](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/date-and-time-types.html),c2.FECHA\_NACIMIENTO,CURDATE()) as edad FROM lu\_cliente c2, fact\_ventas v1, lu\_producto p1 WHERE v1.PRODUCTO\_ID = p1.PRODUCTO\_ID [AND](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/logical-operators.html#operator_and) v1.CLIENTE\_ID = c2.CLIENTE\_ID GROUP BY p1.PRODUCTO\_ID ) as sub01 WHERE sub01.edad BETWEEN 18 [and](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/logical-operators.html#operator_and) 30 GROUP by sub01.edad;

\*31 a 45

[SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) sub01.producto, sub01.edad, sub01.cant FROM ( [SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) [COUNT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/group-by-functions.html#function_count)(p1.PRODUCTO\_ID) as cant, p1.PRODUCTO\_DESC AS producto, TIMESTAMPDIFF([YEAR](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/date-and-time-types.html),c2.FECHA\_NACIMIENTO,CURDATE()) as edad FROM lu\_cliente c2, fact\_ventas v1, lu\_producto p1 WHERE v1.PRODUCTO\_ID = p1.PRODUCTO\_ID [AND](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/logical-operators.html#operator_and) v1.CLIENTE\_ID = c2.CLIENTE\_ID GROUP BY p1.PRODUCTO\_ID ) as sub01 WHERE sub01.edad BETWEEN 31 [and](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/logical-operators.html#operator_and) 45 GROUP by sub01.edad;

\*46 a 60

[SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) sub01.producto, sub01.edad, sub01.cant FROM ( [SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) [COUNT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/group-by-functions.html#function_count)(p1.PRODUCTO\_ID) as cant, p1.PRODUCTO\_DESC AS producto, TIMESTAMPDIFF([YEAR](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/date-and-time-types.html),c2.FECHA\_NACIMIENTO,CURDATE()) as edad FROM lu\_cliente c2, fact\_ventas v1, lu\_producto p1 WHERE v1.PRODUCTO\_ID = p1.PRODUCTO\_ID [AND](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/logical-operators.html#operator_and) v1.CLIENTE\_ID = c2.CLIENTE\_ID GROUP BY p1.PRODUCTO\_ID ) as sub01 WHERE sub01.edad BETWEEN 46 [and](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/logical-operators.html#operator_and) 60 GROUP by sub01.edad;

\*60 a 75

[SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) sub01.producto, sub01.edad, sub01.cant FROM ( [SELECT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/select.html) [COUNT](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/group-by-functions.html#function_count)(p1.PRODUCTO\_ID) as cant, p1.PRODUCTO\_DESC AS producto, TIMESTAMPDIFF([YEAR](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/date-and-time-types.html),c2.FECHA\_NACIMIENTO,CURDATE()) as edad FROM lu\_cliente c2, fact\_ventas v1, lu\_producto p1 WHERE v1.PRODUCTO\_ID = p1.PRODUCTO\_ID [AND](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/logical-operators.html#operator_and) v1.CLIENTE\_ID = c2.CLIENTE\_ID GROUP BY p1.PRODUCTO\_ID ) as sub01 WHERE sub01.edad BETWEEN 60 [and](http://localhost/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/8.0/en/logical-operators.html#operator_and) 75 GROUP by sub01.edad;

*Tarea 6: Análisis*

Utilizar las herramientas necesarias para realizar el análisis requerido en cada caso y de ser necesario crear los algoritmos correspondientes.

Dado un producto, se requiere encontrar el producto con el que más frecuentemente se adquiere. ¿Qué algoritmo de minería de datos utilizaría? ¿Por qué? ¿Puede implementarlo? Piense que, extendido a todos los productos del Data warehouse, brindaría importante información para la organización de la disposición y publicidad de los productos, para establecer recomendaciones a la hora de comprar y para la toma de decisiones acerca de estrategias a aplicar sobre ellos.

* Al tratarse de un producto que se refiere a otro, el algoritmo más óptimo a usar sería el algoritmo de reglas de asociación, debido a la relación que se debe evaluar dentro de un conjunto, es decir, productos.

La realización del algoritmo tendría en cuenta las ocurrencias de ítems por ventas, calculando la frecuencia de compras de un conjunto de ítems en base a un valor mínimo establecido.

*Tarea 7: Presentación.*

Elaborar una memoria del trabajo siguiendo los puntos solicitados, agregar links para exponer y todos los gráficos y páginas necesarios para mejorar la interpretación de los datos.

El trabajo puede ser grupal, la exposición será individual y para todos los compañeros del curso.