INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

CAMPUS QUERÉTARO



Lab_BDA_ETL-3

DBA

Daniel Cu Sánchez - A01703613

ITESM Campus Querétaro Bases de Datos Avanzadas Lab: Ejemplos de ETL Fecha de entrega: martes 26 abril de 2022

Para obtener la calificación del presente laboratorio, es necesario tener la evidencia del desarrollo del mismo. La evidencia será un archivo PDF el cual contendrá las imágenes de pantallas que se van a ir indicando en el desarrollo del laboratorio como ***EVIDENCIA***. El archivo lo suben a canvas en la tarea nombrada "Lab: Éjemplos ETL" El nombre del archivo será: Matricula_LabETL.pdf.

En el presente laboratorio revisaremos varios ejemplos que puedan ayudar con el proceso **ETL** de sus **cubos** del **DWH**.

Parte I. Configurar el ambiente

Para esta parte sigue las instrucciones del profesor

- a) Para facilitar el manejo del DWH ejecuta el script OLTP_EM2022.sql en el esquema de tu matricula. El script creara los objetos (base de datos transaccional). Una vez creados los objetos carga los datos a la tabla VENTAS a partir del archivo datos_ventas.csv
- b) Adicional en la base de datos OLAP (base de datos del repositorio del DWH con la cuenta DWHA0XXXXXXX) ejecutar el script OLAP_EM2022.sql el cual creara los objetos que nos ayudaran en el desarrollo del presente laboratorio.

Parte II. Revisión de vistas, tablas y procedimientos.

Conéctate al esquema **DWHMatricula -OLAP-** y ejecuta las siguientes instrucciones

select count (1) from v_catalogoproductos;

select count (1) from v_catalogoproveedores;

select count (1) from v_catalogosucursales;

select count (1) from v ventas;

EVIDENCIA

racle SQL Developer: DWA01703613 ivo <u>E</u>ditar <u>V</u>er <u>N</u>avegar <u>Ejecutar Origen Equipo</u> <u>H</u>erramientas <u>V</u>entana A<u>y</u>uda the DWA01703613.sql × □ OLTP_EM2022.sql × □ Página de bienvenida × the DWA01703613 kiones · 🚱 🔬 🕰 🖶 acle conexiones Hoja de Trabajo Generador de Consultas A01703613 select count (1) from v_catalogoproductos; Africa select count (1) from v_catalogoproveedores; select count (1) from v_catalogosucursales; select count (1) from v_ventas; America Asia DanielCu DWA01703613 Tablas (Filtrado) Wistas Vistas indices Paquetes Procedimientos Funciones dos los Informes Informes de Diccionario de Datos Informes Definidos por el Usuario <u>a</u> 🖶 nexiones Salida de Script × A01703613 🏓 🤌 🖥 🚨 📘 | Tarea terminada en 0.096 segundos 12 ción COUNT (1) ⊕ □ □ □ oyectos: (Seleccione el repositorio de migración COUNT (1) COUNT (1)

Revisa la estructura de las vistas.

Considerando el siguiente esquema estrella para un cubo de un datawarehouse:



Para la dimensión de "Productos", vamos a tener los siguientes atributos:

- Descripción del producto
- Nombre del proveedor
- Código del producto (base de datos transaccional)

Para la dimensión "Ubicación", vamos a tener los siguientes atributos:

- Region
- Código de sucursal (base de datos transaccional)

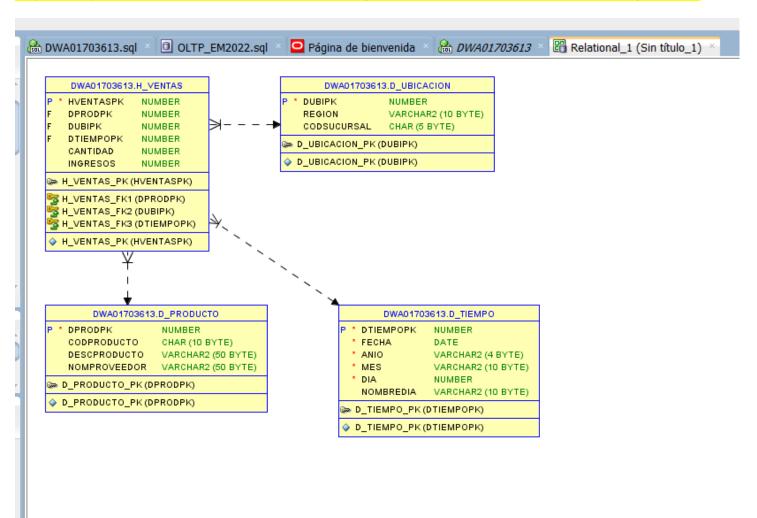
Para la dimensión "Tiempo", vamos a tener los siguientes atributos:

- Fecha
- Año
- Mes
- Día
- Nombre del día

Revisa las estructuras de las tablas **D_PRODUCTOS**, **D_TIEMPO**, **D_UBICACION** y **H_VENTAS**. De la misma manera revisa los **constraints** entre cada una de las tablas.

EVIDENCIA

Captura la pantalla con la estructura de las tablas y las relaciones entre las dimensiones y hechos



Parte III. Implementación de procedimiento de ETL para tablas de dimensiones.

Para mantener **actualizadas** las **tablas** de **dimensiones** es necesario extraer la información de las bases de datos transaccionales. En nuestro caso para facilitar la extracción creamos las vistas que nos ayudaran en el proceso. Solamente vamos a insertar los registros "**nuevos**" de nuestra base de datos transaccional. Para esta práctica consideramos que no debemos borrar ningún registro del **DWH** aun cuando se borren en la base de datos transaccional, esto para mantener **información** "**histórica**" en el **DWH**.

Para la **dimensión producto** actualiza el procedimiento **ACTUALIZA_PRODUCTO** de la siguiente manera (recuerda estar ya conectado al esquema dwhMatricula):

```
create or replace PROCEDURE ACTUALIZA_PRODUCTO AS

BEGIN

insert into d_producto
select seq_d_producto.nextval, codproducto, descripcion, razonsocial
from v_catalogoproductos cp, v_catalogoproveedores cv
where cp.rfcproveedor = cv.rfcproveedor
and cp.codproducto not in (select codproducto from d_producto);
commit;

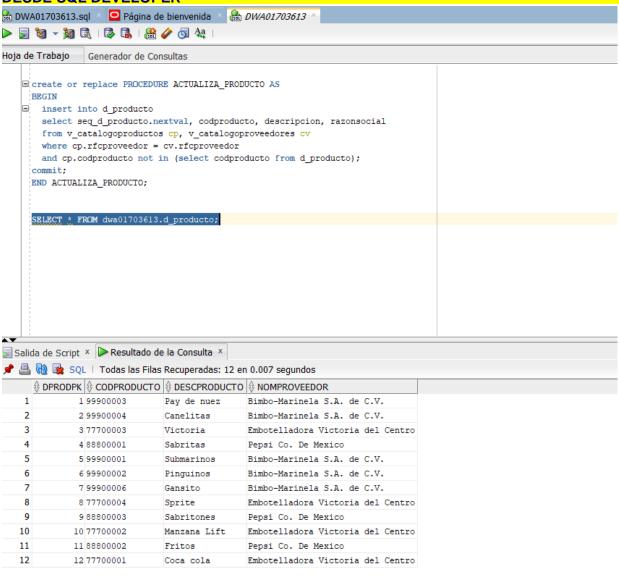
END ACTUALIZA PRODUCTO;
```

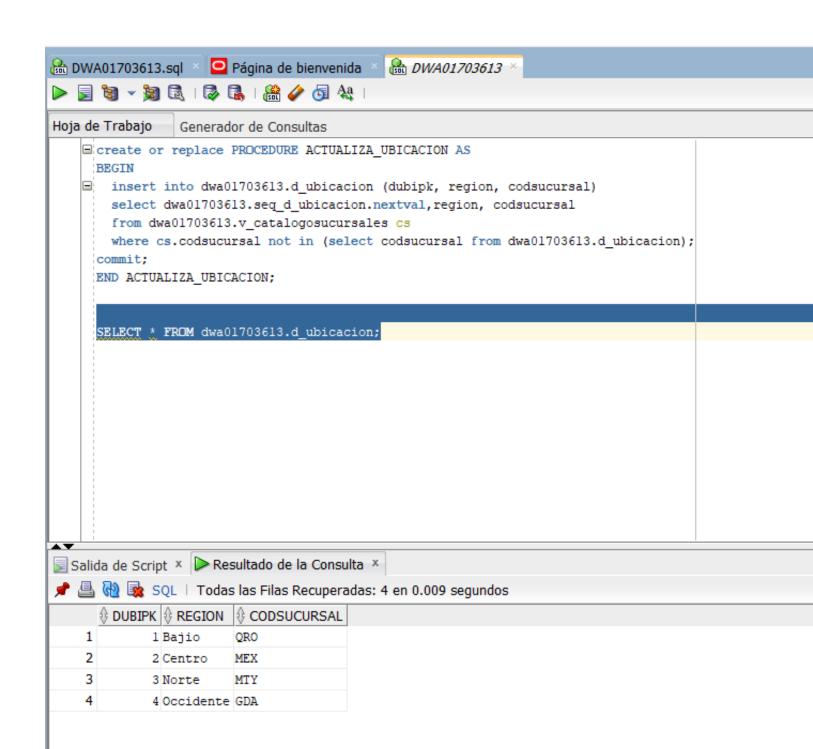
Para la dimensión ubicación actualiza el procedimiento ACTUALIZA_UBICACION a partir de la siguiente consulta:

```
select seq_d_ubicacion.nextval, region, codsucursal
from v_catalogosucursales cs
where cs.codsucursal not in (select codsucursal from d ubicacion);
```

Ejecuta los **dos procedimientos** que acabas de modificar y en este punto ya tenemos pobladas las dos tablas de dimensiones del cubo.

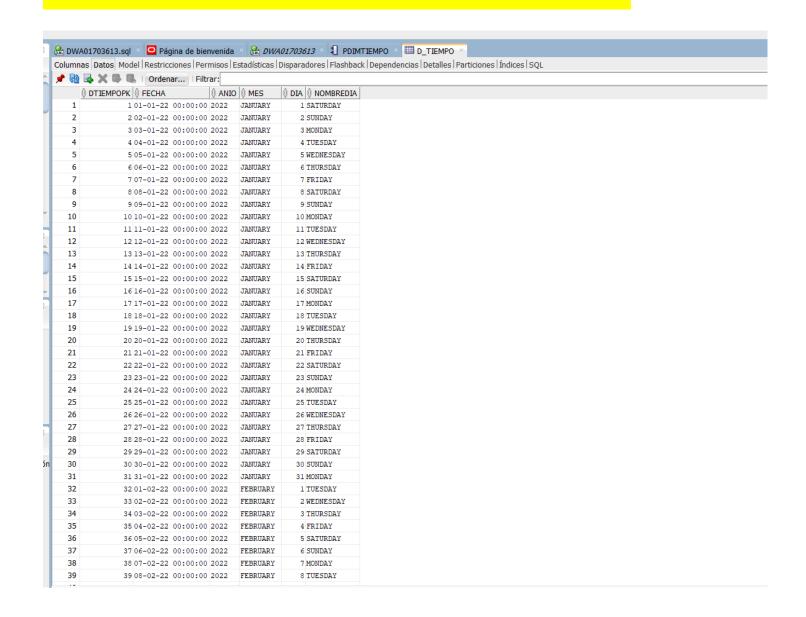
EVIDENCIA REALIZA LA CONSULTA A CADA UNA DE LAS TABLAS QUE SE ACABAN DE POBLAR Y TOMA LA CAPTURA DE LA MISMA – CONSULTA Y DATOS Y/O PESTAÑA DE DATOS DESDE SQL DEVELOPER-





Para la dimensión tiempo es necesario ejecutar el procedimiento PDIMTIEMPO para poblar con datos entre las fechas 1 de enero de 2022 y 21 de abril de 2022.

EVIDENCIA REALIZA LA CONSULTA A LA TABLA S QUE SE ACABAN DE POBLAR Y TOMA LA CAPTURA DE LA MISMA – CONSULTA Y DATOS Y/O PESTAÑA DE DATOS DESDE SQL DEVELOPER-



Parte IV Implementación de procedimiento de ETL para la tabla de hechos.

Para la implementación de la **tabla** de **hechos** vamos a tomar las siguientes consideraciones:

- Debemos especificar la fecha de inicio y la fecha de fin de la extracción de datos. Esto puede ser muy útil para una carga inicial de datos y después para cargas incrementales (reducción de cargas de trabajo tanto en las bases de datos transaccionales como la del DWH), por ejemplo, cargas diarias o semanales.
- Para evitar la carga de datos "duplicados" o errores en los constraints para periodos de tiempo ya capturados, los procedimientos de cargas de hechos deben poder actualizar y/o sustituir la información ya existente. Una manera fácil de realizar esto es borrando antes de la inserción los registros que correspondan al periodo de tiempo que se va a actualizar. En nuestro caso el procedimiento de carga debe iniciar con el borrado de los registros que corresponden al periodo del tiempo a insertar.
- Los indicadores de cantidad e ingresos se calculan de la siguiente manera o cantidad sum(ventas.cantidad)
 - o ingresos o sum(ventas.cantidad*ventas.precio)
- Todas las dimensiones deben agruparse juntas, considerando el tiempo, producto y ubicación.
- Para el caso de la dimensión tiempo se debe usar la **función trunc** de la fecha para que sea fácil la comparación de las fechas.

Teniendo en cuenta las consideraciones anteriores, **escribe** una **consulta** que obtenga el **identificador** del **producto**, **ubicación** y **tiempo** de las **dimensiones** y se calcule los **indicadores** a partir de la **vista** de **ventas**. Usa la **función** de **agrupación** para que puedas realizar la consulta. Las condiciones deben cumplir con lo siguiente: o Para la dimensión producto, usar el correspondiente código de producto transaccional con su correspondiente registro en la vista de ventas. o Para la dimensión ubicación, usar el correspondiente código de ubicación transaccional con su correspondiente registro en la vista de ventas.

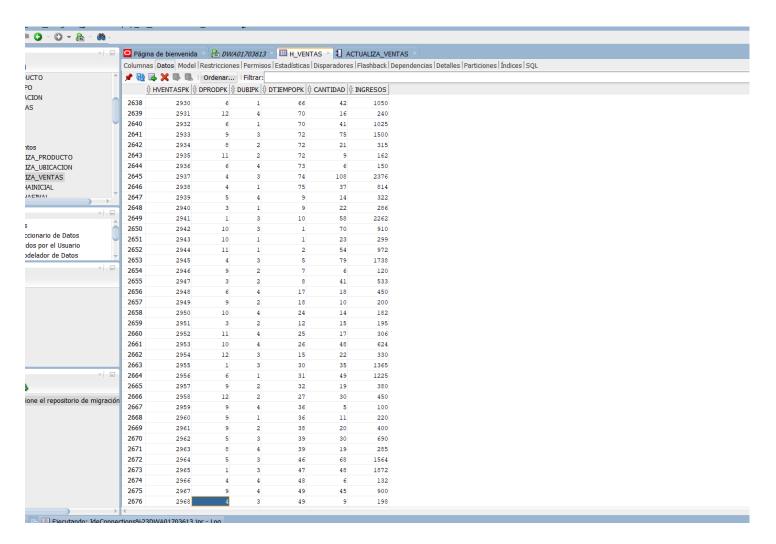
Para la dimensión tiempo usa la función **between fechainicio and fechafinal** para obtener los registros de la vista de ventas y realizar el **join** con la función **trunc** a la correspondiente columna de la dimensión del tiempo.

Una vez que tengas la consulta, ajusta el procedimiento **ACTUALIZA VENTAS** con la siguiente estructura:

```
create or replace
PROCEDURE ACTUALIZA_VENTAS
  FECHAINICIAL IN DATE
 , FECHAFINAL IN DATE
-) AS
  vFechaInicial date;
  vFechaFinal date;
  vdProPk number;
vdTiempoPk number;
  vdUbiPk number;
  v_cantidad number;
  v_ingresos number;
  cursor c tiempo is
  select dtiempopk
  from d tiempo
  where fecha between vFechaInicial and vFechaFinal;
  cursor c ventas is
  -- código de la consulta desarrollada_
BEGIN
  vFechaInicial := FECHAINICIAL;
  vFechaFinal := FECHAFINAL;
  open c tiempo;
  LOOP
    fetch c_tiempo into vdTiempoPk;
    exit when c_tiempo%NOTFOUND;
    delete from h_ventas where dtiempopk=vdTiempoPk;
    commit;
 END LOOP;
  close c_tiempo;
  open c_ventas;
 LOOP
    fetch c_ventas into vdProPk, vdUbiPk, vdTiempoPk, v_cantidad, v_ingresos;
    exit when c ventas%NOTFOUND;
    insert into h_ventas (hventaspk,dprodpk,dubipk,dtiempopk,cantidad,ingresos)
    values (seq h_ventas.nextval, vdProPk, vdUbiPk, vdTiempoPk, v_cantidad, v_ingresos);
    commit;
  END LOOP;
  close c ventas;
END ACTUALIZA VENTAS;
```

Ejecuta el procedimiento ACTUALIZA_VENTAS para poblar con datos entre las fechas 1 de enero de 2022 y 18 de marzo de 2022.

EVIDENCIA CAPTURA DE PANTALLA EN DONDE SE MUESTRE EL TOTAL DE REGISTROS DE LA TABLA DE HECHOS



```
create or replace PROCEDURE ACTUALIZA_VENTAS
(
   FECHAINICIAL IN DATE,
   FECHAFINAL IN DATE
)
AS
VFechaInicial date;
vFechaFinal date;
vdProPk number;
vdTiempoPk number;
vdTiempoPk number;
v_cantidad number;
v_cantidad number;
v_ingresos number;
cursor c_tiempo is
select dtiempopk
from dwa01703613.d_tiempo
where fecha BETWEEN vFechaInicial and vFechaFinal;
```

```
cursor c_ventas is
select DPO.dprodpk, DUB.dubipk, TI.dtiempopk, SUM(VE.cantidad), SUM(VE.cantidad * VE.precio)
from a01703613.ventas VE, dwa01703613.d_producto DPO, dwa01703613.d_tiempo TI,
dwa01703613.d_ubicacion DUB
where VE.codproducto = DPO.codproducto
and VE.codsucursal = DUB.codsucursal
and TRUNC (VE.fechahora) = TI.fecha
group by DPO.dprodpk, DUB.dubipk, TI.dtiempopk;
BEGIN
  vFechaInicial := FECHAINICIAL;
  vFechaFinal := FECHAFINAL;
  open c_tiempo;
  L00P
    fetch c_tiempo into vdTiempoPk;
    exit when c tiempo%NOTFOUND;
    delete from dwa01703613.h_ventas where dtiempopk=vdTiempoPk;
    commit;
  END LOOP;
  close c_tiempo;
  open c_ventas;
  L00P
    fetch c_ventas into vdProPk, vdUbiPk, vdTiempoPk, v_cantidad, v_ingresos;
    exit when c_ventas%NOTFOUND;
    insert into dwa01703613.h ventas (hventaspk,dprodpk,dubipk,dtiempopk,cantidad,ingresos)
    values (dwa01703613.seq_h_ventas.nextval, vdProPk, vdUbiPk, vdTiempoPk, v_cantidad,
v_ingresos);
    commit;
  END LOOP;
  close c_ventas;
END ACTUALIZA VENTAS;
```

Parte V Consultas

La siguiente consulta nos da como resultado las ventas por producto acumulado por mes desde el 1 de enero hasta el 1 de abril del 2022. Impleméntala en una vista que se llame ventaProductoMes

```
select descproducto as "PRODUCTO",
sum(cantidad) as "CANTIDAD VENDIDA",
MES
from d_producto dp,
h_ventas hv,
d_tiempo dt
where dp.dprodpk = hv.dprodpk
and dt.dtiempopk = hv.dtiempopk
and dt.fecha between to_date ('01-01-2022','DD-MM-YYYY') and to_date('01-04-2022','DD-MM-YYYY')
group by descproducto, MEs
order by 1, to_char(to_date(MES,'Month'),'MM') asc
```

El resultado de la consulta les debe dar algo como esto:

♦ PRODUCTO	
¹Coca cola	3291 JANUARY
² Coca cola	2751 FEBRUARY
3 Coca cola	1899MARCH
⁴Fritos	3353 JANUARY
5 Fritos	2912 FEBRUARY
6 Fritos	1897MARCH
⁷ Manzana Lift	3500 JANUARY
8 Manzana Lift	
⁹ Manzana Lift	1986MARCH
10 Pav de nuez	3416 JANUARY
¹¹ Pay de nuez	2968 FEBRUARY
12 Pay de nuez	1736MARCH
13 Pinquinos	3154 JANUARY
14 Pinquinos	3188 FEBRUARY
15 Pinquinos	1698MARCH
¹⁶ Sabritas	3297 JANUARY
¹⁷ Sabritas	2973 FEBRUARY
18 Sabritas	1854MARCH
19 Sabritones	3009 JANUARY
20 Sabritones	3203 FEBRUARY
²¹ Sabritones	1977 MARCH
²² Sprite	3623 JANUARY
²³ Sprite	2679 FEBRUARY
²⁴ Sprite	1610MARCH
25 Submarinos	3156 JANUARY
26 Submarinos	2982 FEBRUARY
²⁷ Submarinos	1864MARCH
²⁸ Victoria	3372 JANUARY
²⁹ Victoria	2827 FEBRUARY
30 Victoria	1913MARCH

Siguiendo el mismo ejemplo anterior implementen las vistas siguientes:

- Ventas de cada uno de los productos para todos los viernes del mes de FEBRERO
- Ventas de "Pay de Nuez" en cada sucursal
- · Ventas de "Pinguinos" en cada sucursal, pero solamente los miércoles
- Ventas de TODOS los productos por sucursal y mes

EVIDENCIA

CAPTURA DE PANTALLA EN DONDE SE MUESTRE LA CONSULTA Y EL RESULTADO DE LAS MISMAS – 5 REPORTES – ANTERIORES.

```
-- Ventas de cada uno de los productos para todos los viernes del mes de FEBRERO

select descproducto as "PRODUCTO", sum(cantidad) as "CANTIDAD VENDIDA"

from d_producto dp,h_ventas hv,d_tiempo dt,d_ubicacion du

where dp.dprodpk = hv.dprodpk and dt.dtiempopk =hv.dtiempopk and du.dubipk=hv.dubipk

and dt.fecha between to_date ('01-02-2022','DD-MM-YYYY') and to_date('28-02-2022','DD-MM-YYYY')

and dt.nombredia ='FRIDAY'

group by descproducto

order by 1 asc;
```

PRODUCTO	CANTIDAD VENDIDA
Coca cola	455
Fritos	432
Manzana Lift	388
Pay de nuez	414
Pinguinos	434
Sabritas	363
Sabritones	525
Sprite	413
Submarinos	410
Victoria	429
10 filas seleccionadas.	

```
-- Ventas de "Pay de Nuez" en cada sucursal

select descproducto as "PRODUCTO",du.codsucursal,sum(cantidad) as "CANTIDAD VENDIDA"

from d_producto dp,h_ventas hv,d_tiempo dt,d_ubicacion du

where dp.dprodpk = hv.dprodpk and dt.dtiempopk =hv.dtiempopk and du.dubipk=hv.dubipk

and dt.fecha between to_date ('01-01-2022','DD-MM-YYYY') and to_date('01-04-2022','DD-MM-YYYY')

and dp.descproducto ='Pay de nuez'

group by descproducto, du.codsucursal

order by 1 asc;
```

	PROI	DUC:	0	CODSU	CANTIDAD	VENDIDA
	Pay	de	nuez	QRO		1940
ı	Pay	de	nuez	MEX		1052
l	Pay	de	nuez	MTY		3615
	Pay	de	nuez	GDA		1513
ı						

```
--Ventas de "Pinguinos" en cada sucursal, pero solamente los miércoles

select descproducto as "PRODUCTO",du.codsucursal,sum(cantidad) as "CANTIDAD VENDIDA"

from d_producto dp,h_ventas hv,d_tiempo dt,d_ubicacion du

where dp.dprodpk = hv.dprodpk and dt.dtiempopk =hv.dtiempopk and du.dubipk=hv.dubipk

and dt.fecha between to_date ('01-01-2022','DD-MM-YYYY') and to_date('01-04-2022','DD-MM-

YYYY')

and dp.descproducto ='Pinguinos' and dt.nombredia ='WEDNESDAY'

group by descproducto, du.codsucursal

order by 1 asc;
```

П				
	PRODUCTO	CODSU	CANTIDAD	VENDIDA
	Pinguinos	MEX		270
	Pinguinos	QRO		266
	Pinguinos	MTY		561
	Pinguinos	GDA		289
- 11				

--Ventas de TODOS Los productos por sucursal y mes
select descproducto as "PRODUCTO",du.codsucursal,sum(cantidad) as "CANTIDAD VENDIDA"
from d_producto dp,h_ventas hv,d_tiempo dt,d_ubicacion du
where dp.dprodpk = hv.dprodpk and dt.dtiempopk =hv.dtiempopk and du.dubipk=hv.dubipk
and dt.fecha between to_date ('01-01-2022','DD-MM-YYYY') and to_date('01-04-2022','DD-MMYYYY')
group by descproducto, du.codsucursal
order by 1 asc;

PRODUCTO	CODSU	CANTIDAD VENDIDA
Coca cola	GDA	1659
Coca cola	MEX	1087
Coca cola	MTY	3415
Coca cola	QRO	1780
Fritos	GDA	1581
Fritos	MEX	1082
Fritos	MTY	3406
Fritos	QRO	2093
Manzana Lift	GDA	1656
Manzana Lift	MEX	1062
Manzana Lift	MTY	3560
Manzana Lift	QRO	2009
Pay de nuez	GDA	1513
Pay de nuez	MEX	1052
Pay de nuez	MTY	3615
Pay de nuez	QRO	1940
Pinguinos	GDA	1504
Pinguinos	MEX	1246
Pinguinos	MTY	3090
Pinguinos	QRO	2200
Sabritas	GDA	1414
Sabritas	MEX	1022
PRODUCTO	CODSU	CANTIDAD VENDIDA
Sabritas	MTY	3607
Sabritas	QRO	2081
Sabritones	GDA	1572
Sabritones	MEX	1051
Sabritones	MTY	3417
Sabritones	QRO	2149
Sprite	GDA	1556