



The Drinking Company

Informe DataMart Ventas

**López Ceratto Julieta
Slepoy, David
Rosito, Valentín**

ÍNDICE

CONTENIDO DEL INFORME

Información de la empresa	1
Fuentes de datos	2
Diseño Data Mart Ventas	3
ETL: Cargas iniciales	9
ETL: versionado	3
Reportes	14
	15

01 | THE DRINKING CO.



The Drinking Company es una empresa con trayectoria de 4 años, dedicada a la producción de bebidas y su comercialización tanto en el mercado minorista como mayorista.

02 FUENTES DE DATOS

TDCHISORYSALES

Base de datos SQLSERVER 2000 de ventas históricas (history sales);
Contiene ventas hasta el año 2008

SALES

Base de datos MYSQL, contiene ventas a partir del año 2009, como así también los descuentos y precios aplicados históricos.

ARCHIVOS CSV

Bajo este formato, se encuentran los archivos:

- Products : contiene información de los productos (nombre, envase, cm3/litros, entre otros).
- Regions: contiene las regiones con las ciudades y sus códigos postales.
- Stock: contiene información acerca del stock disponible para un producto en cierta fecha, de forma histórica.

ARCHIVOS EXCEL

Bajo este formato, se encuentran los archivos:

- Employee: contiene información acerca de los empleados (datos personales, fecha de contratación, categoría, etc).
- Holidays: contiene información sobre los días vacacionales.

ARCHIVOS XML

Bajo este formato, se encuentran los archivos:

- Customer_R.
- Customer_W.

Detallan información tanto de los clientes minoristas (R) como mayoristas (W)

03 | DISEÑO DATAMART VENTAS

El desarrollo del diseño de dicho Data Mart, constó del análisis de los datos disponibles, el hecho a analizar, las consultas precisadas por la empresa, así como también de los recursos disponibles de la misma.

A continuación se detallan las dimensiones y los atributos de las mismas, así como también del hecho.

F_VENTAS

F_Ventas es la tabla de hechos. La empresa busca analizar, al fin y al cabo, las ventas producidas de forma histórica. Esta tabla está conformada de la siguiente forma:

ID_Venta: es el id del hecho en el DataMart.

ID_Sistema_Origen: ID de la base de datos del sistema de origen.

Num_Comprobante_origen: numero de la factura en el sistema de origen.

Fecha: fecha del hecho.

ID_Cliente: Id del cliente en el sistema de origen.

ID_Emppleado: id del empleado en el sistema de origen.

Sistema_Origen: sistema que indica el origen de los datos para dicha fila.

ID_Descuento: en la D_Descuento.

F_Venta *
ID_Venta
ID_Sistema_Origen
Num_Comprobante_origen
Sistema_Origen
Fecha
ID_Cliente
ID_Emppleado
ID_Producto
Sucursal
Region
Fecha_Sistema
H_Venta
Cant_Producto
Precio_Unitario
Importe_Antes_Descuento
ID_Descuento
Importe_Descuento
Importe_Final
Cm3_Totales
Litros_Totales

ID_Producto: id del producto en el sistema de origen.

Sucursal: nombre de la sucursal donde ocurrió el hecho.

Región: región donde se produjo la venta.

Fecha_Sistema: fecha detallada a nivel segundos, es la fecha del sistema.

H_Venta: hora de la venta.

Cant_Producto: cantidad de unidades del producto.

Precio_Unitario: del producto.

Importe_Antes_Descuento: importe bruto.

Importe_Descuento: importe correspondiente al descuento sobre el importe antes de descuento.

Importe_Final: importe final abonado por el cliente luego de los descuentos aplicados.

Cm3_Totales y Litros_Totales: cantidad en esa unidad total del producto.

D_EMPLEADO

Esta tabla representa la Dimensión Empleado, como lo indica su nombre, tiene la información histórica y versionada de los empleados de la empresa.

ID_Empleado: Es el id que identifica un empleado y su versión dentro del DataMart.

Legajo empleado: legajo / id empleado en su origen.

Nombre: Nombre del empleado.

Apellido: Apellido del empleado.

Género: género del empleado.

D Empleado	
ID_Empleado	
Legajo_Empleado	
Nombre	
Apellido	
Género	
F_ingreso	
F_nacimiento	
Nivel_educativo	
ID_Categoría_Empleado	
Sistema_Origen	
fecha_desde_version	
fecha_hasta_version	
version	

F_ingreso: fecha en la que el empleado comenzó a trabajar en la empresa.

F_nacimiento: fecha de nacimiento del empleado.

Nivel educativo: máximo grado educativo alcanzado por el empleado.

ID_Categoría_Empleado: Id correspondiente a la categoría del empleado en la Dimensión Categoría Empelado.

Sistema_Origen: sistema que indica el origen de los datos para dicha fila.

fecha_desde_version: indica la fecha a partir de la cual la versión del empleado está/estuvo vigente.

fecha_hasta_version: indica la fecha hasta la cual la versión del empleado está/estuvo vigente.

D_CATEGORIA_EMPLEADO

Esta tabla representa la Dimensión Categoría Empleado, como lo indica su nombre, tiene la información de las categorías de los empleados dentro de la empresa.

D Categoría Empleado	
ID_Categoría_Empleado	
Categoría_Empleado	

ID_Categoría_Empleado: id que identifica la categoría del empleado dentro del DataMart

Categoría_Empleado: nombre de la categoría de empleado de la empresa.

D_CLIENTE

Esta tabla representa la Dimensión Cliente, como lo indica su nombre, tiene la información de los clientes de la empresa; tanto minoristas como mayoristas.

ID_Cliente: id identificadorio del cliente dentro del DataMart.

ID_Cliente_Origen: Id del cliente proveniente del origen de los datos.

Nombre: nombre del cliente.

Apellido: apellido del cliente.

D Cliente	
?	ID_Cliente
	ID_Cliente_Origen
	Nombre
	Apellido
	Fecha_Nacimiento
	Tipo_Cliente
	Zip_Code
	Ciudad
	Estado
	Sistema_Origen

Fecha_Nacimiento: fecha de nacimiento del cliente.

Tipo_Cliente: minorista / mayorista.

Zip_Code: código postal correspondiente al lugar de residencia del cliente.

Ciudad: ciudad de residencia del cliente.

Estado: estado de residencia del cliente.

Sistema_Origen: indica el sistema de procedencia de dichos datos.

D_PRODUCTO.

Esta tabla representa la Dimensión Producto, como lo indica su nombre, tiene la información de los productos de la empresa, de manera versionada.

ID_Producto: id identificadorio del producto y versión dentro del DataMart.

Codigo_Producto: código del producto proveniente del origen de los datos.

Producto: nombre del producto.

Categoría: categoría del producto.

D Producto	
?	ID_Producto
	Codigo_Producto
	Producto
	Categoría
	Presentacion
	Cm3
	Sistema_Origen
	Precio_Producto
	Fecha_desde_producto
	Fecha_hasta_producto

Presentacion: envase del producto.

Cm3: cantidad de Cm³ del envase del producto.

Sistema_Origen: sistema de procedencia de los datos.

Precio_Producto: precio del producto para dicha versión.

Fecha_desde_producto: fecha a partir de la cual está/estuvo vigente dicha versión del producto.

Fecha_hasta_producto: fecha hasta la cual está/estuvo vigente dicha versión del producto.

D_DESCUENTO

Esta dimensión representa los descuentos de la empresa históricos.

ID_Descuento: identifica el descuento dentro del DataMart.

ID_Descuento_Origen: id del descuento en el sistema de origen.

Sistema_Origen: indica el sistema de procedencia de los datos.

Monto_Minimo: monto de la factura para que el descuento sea aplicable.

D Descuento	
ID_Descuento	
ID_Descuento_Origen	
Sistema_Origen	
F_desde	
F_hasta	
Duracion_Descuento	
Monto_Minimo	
[Descuento (%)]	

F_desde: fecha a partir de la cual está/estuvo vigente el descuento.

F_hasta: fecha hasta la cual está/estuvo vigente el descuento.

Duración_Descuento: duración en días del descuento.

Descuento (%): valor en % del descuento aplicable sobre la factura.

D_GEOGRAFÍA.

Esta dimensión representa la forma en la que la empresa distribuye la geografía de las ventas. La distribuye mediante regiones principalmente; a cada una de ellas hay asociadas estados y ciudades con su respectivo zipcode.

Zip_Code: es el código postal, al ser único, es a su vez id de la dimensión dentro del DataMart.

D Geografia	
Zip_Code	
Ciudad	
Estado	
Region	

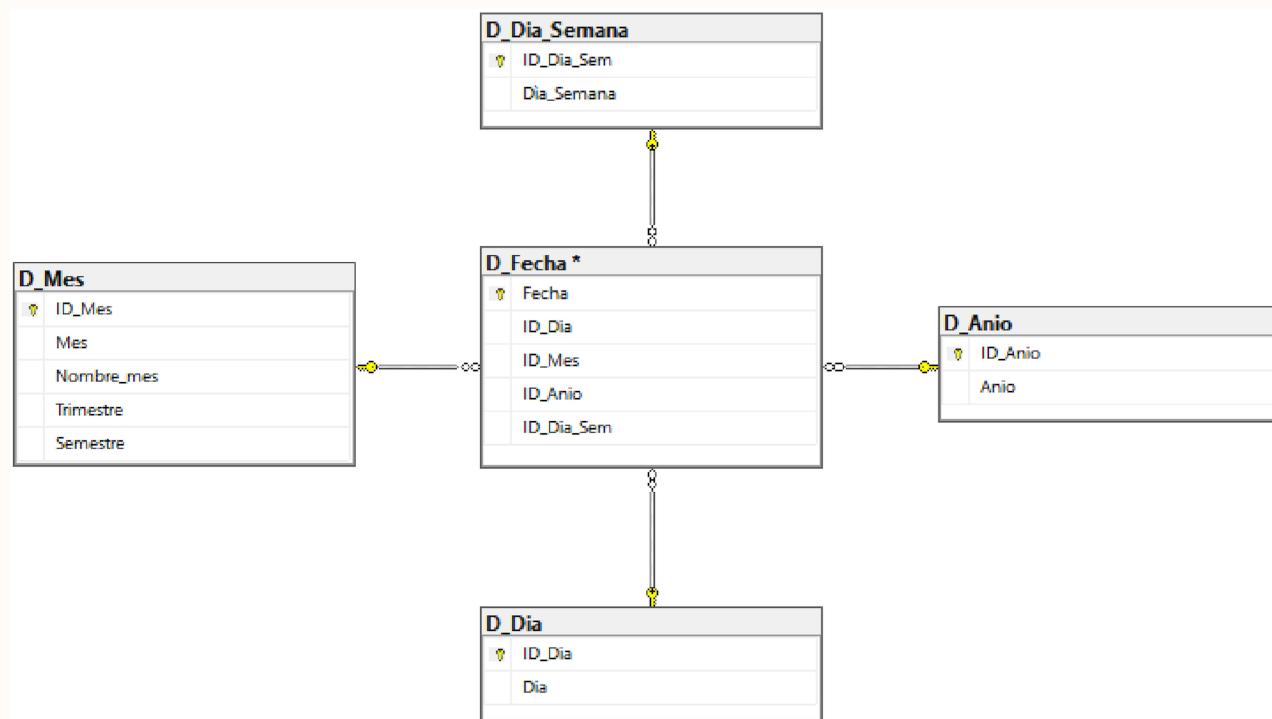
Ciudad: ciudad.

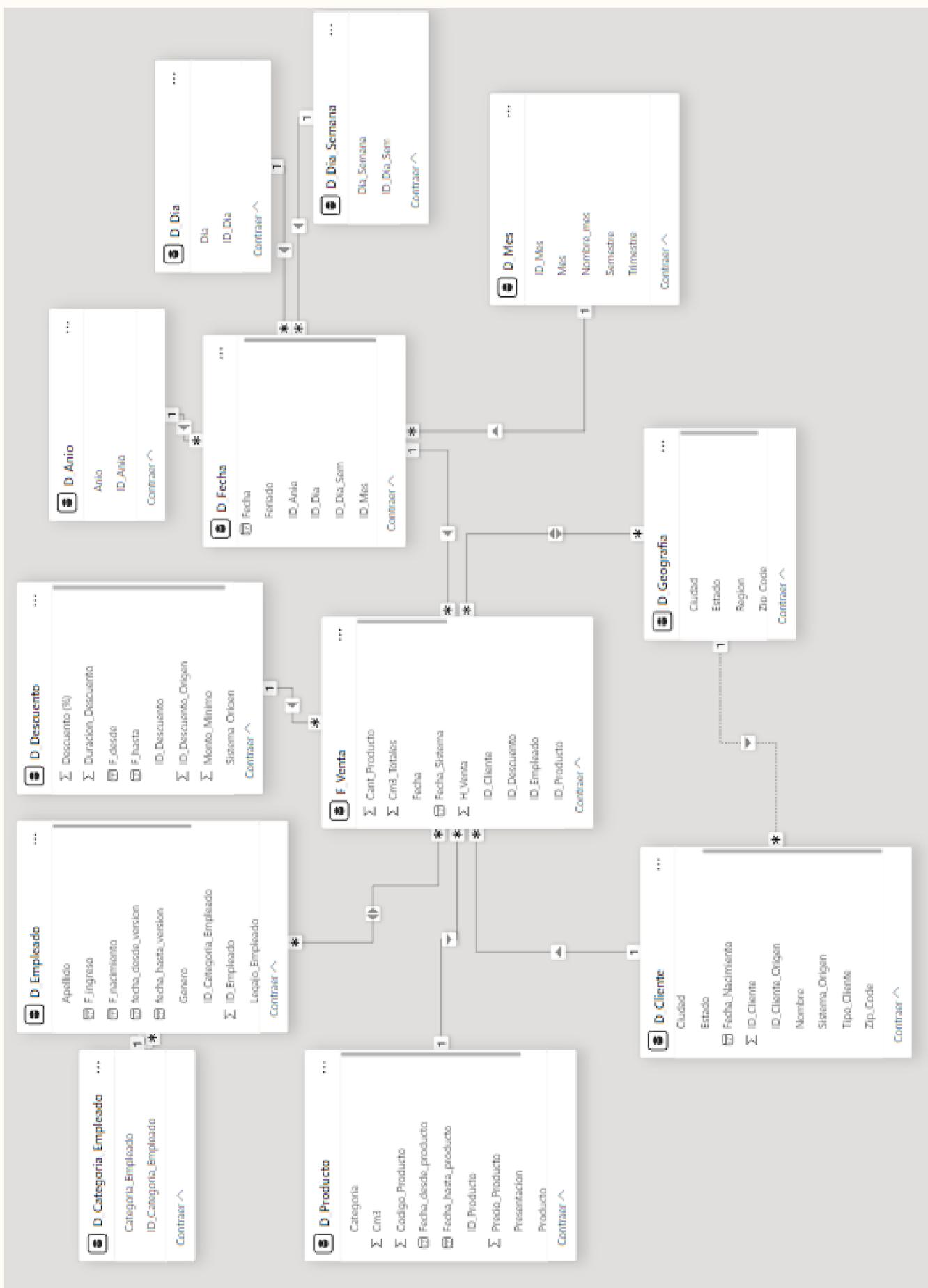
Estado: estado.

Región: región.

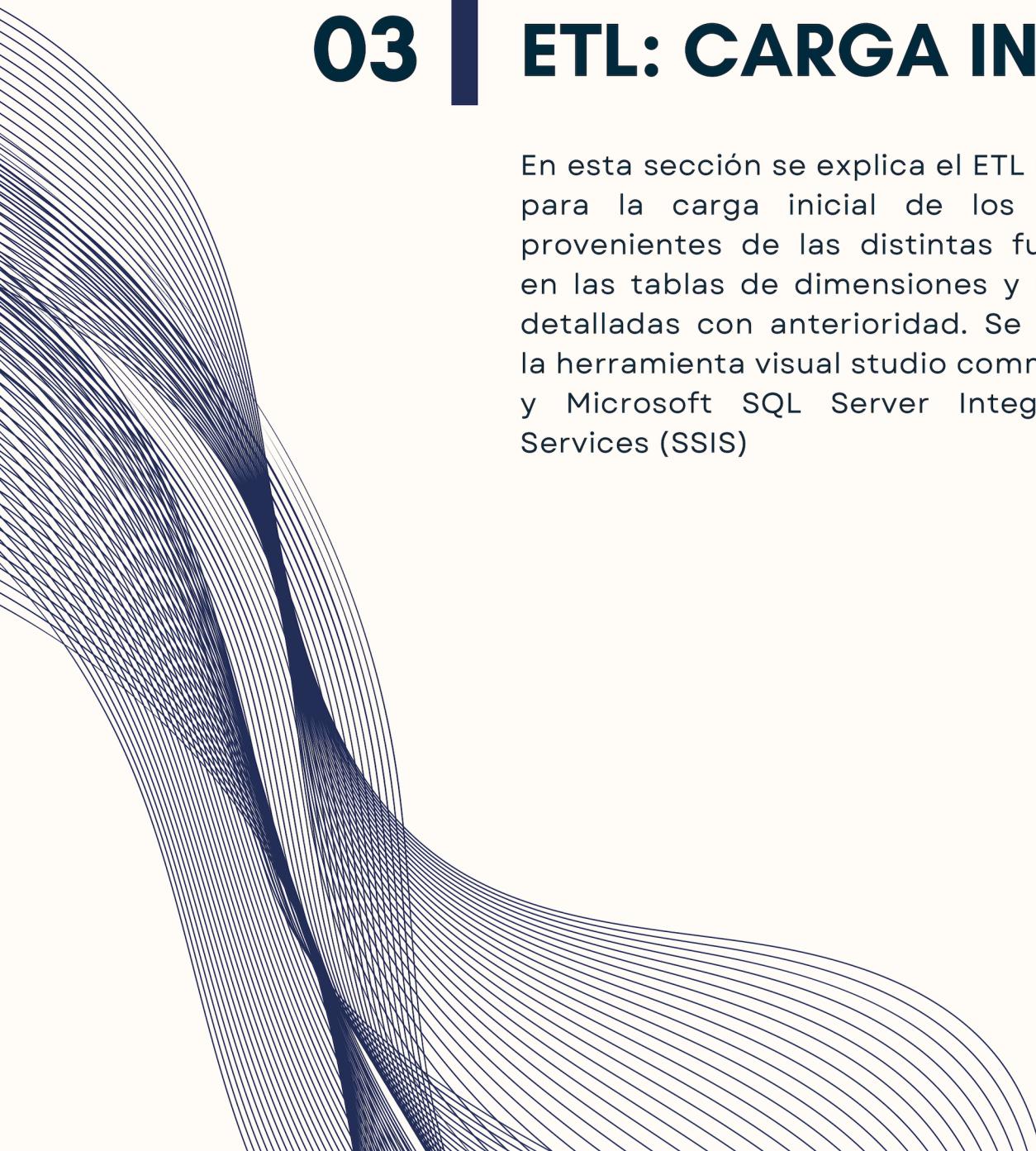
ESTRUCTURA DIMENSIONAL TIEMPO

Esta estructura dimensional representa el tiempo, el cual está compuesto por las siguientes dimensiones.





03 | ETL: CARGA INICIAL



En esta sección se explica el ETL hecho para la carga inicial de los datos provenientes de las distintas fuentes en las tablas de dimensiones y hecho detalladas con anterioridad. Se utilizó la herramienta visual studio community y Microsoft SQL Server Integration Services (SSIS)

ETAPAS

TRANSFORM

Transformación de los datos y agregación de datos nuevos.

EXTRACT

Extracción de los datos de las distintas fuentes de origen en sus diversos formatos.

LOAD

Carga de los datos al DataMart.

CONSIDERACIONES EXTRACT

Los archivos extraídos en cada tarea del ETL están con su correspondiente nombre en las mismas, así como también se puede observar la forma en la que se extraen. Se extraen los datos de 2 formas principalmente. Cargando una conexión desde el propio SSIS, o bien leyéndolos desde el área de stagin.

STAGIN

Se utiliza stagin para casi todas las extracciones de los datos, para luego cargarlos a las dimensiones. Esto se debe a que es más sencillo realizar consultas, agregar columnas, leer los datos, transformarlos y demás mediante lenguaje SQL.

CONSIDERACIONES TRANSFORM

El tratamiento de datos se realizó a fin de que la información almacenada sea consistente, integrada y homogénea.

A continuación se detallan ciertas decisiones tomadas en cuanto a esto.

FECHA HASTA = ‘3000-31-12’

Siempre que como valor de fecha hasta aparezca la fecha ‘3000-12-31’, representa que esa versión, producto, descuento, etc., no tiene una fecha de finalización prevista o bien, sigue vigente.

SISTEMA ORIGEN.

Se agregaron columnas con los sistemas de origen de los datos para que de esta forma sean trazables; es decir, se pueda identificar de dónde se obtuvieron los mismos de forma fácil.

LEGAJOS / ID DE ORIGEN.

Se mantienen los id / legajos de los sistemas de origen ya que permite identificar de manera sencilla el dato en el sistema de origen y entender el por qué de su valor o si hay algún inconveniente, poder encontrarlo en dicho sistema de forma rápida.

ACCIONES EN EL SSIS.

Principalmente en el SSIS se realizaron las acciones de transformación de datos para que sean compatibles con el DataMart, el cual está hecho en SQL Server; o bien acciones de agregado de columnas.

La acción más destacable en este punto es el control y manejo de errores (por ejemplo, validando campos null en los códigos / legajos / id's de origen).

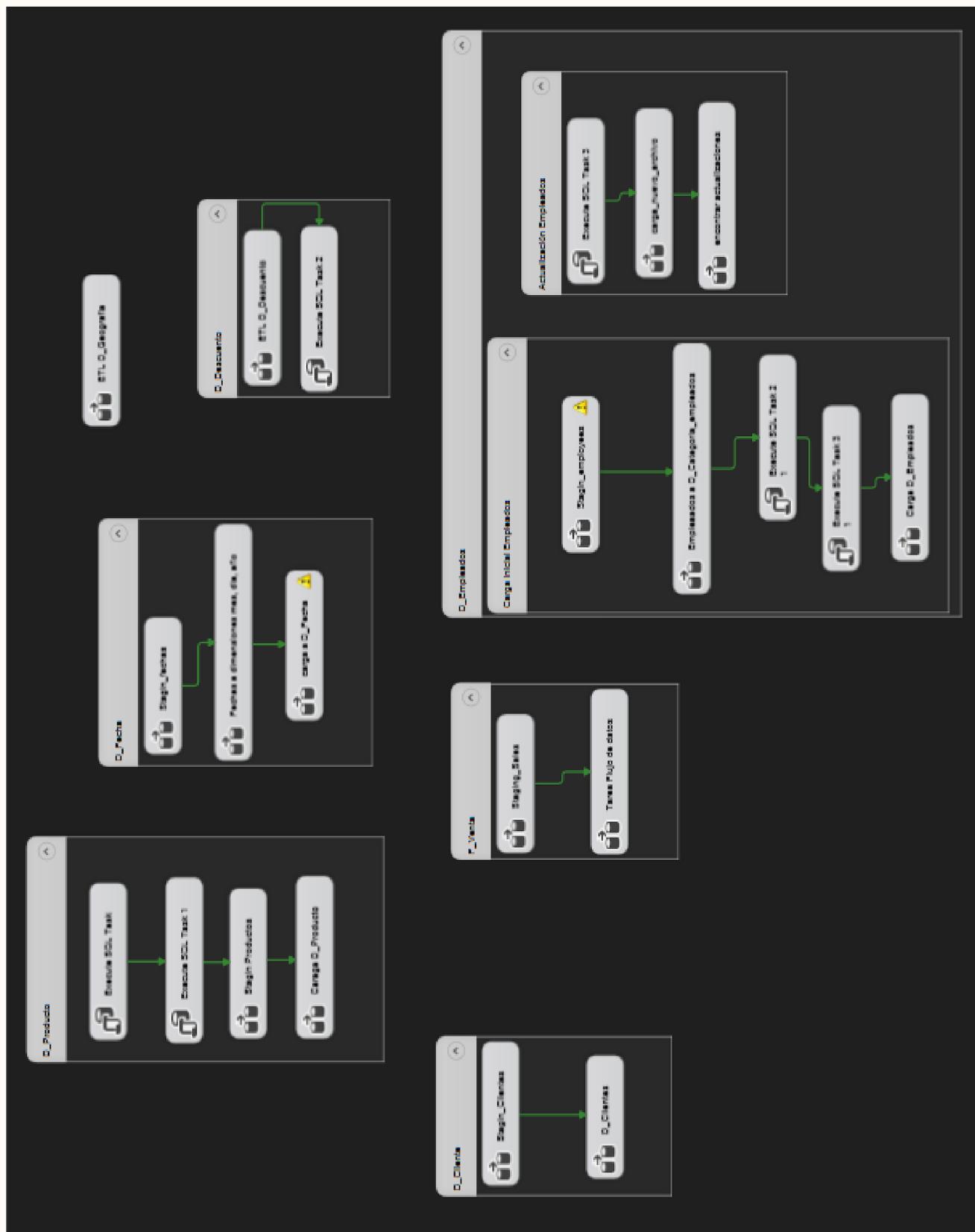
| CONSIDERACIONES LOAD

La carga de datos en el DataMart para cada dimensión y para la tabla de hecho, se realizó en su mayoría a partir de un stagin previo; esto se debe a que los datos cuando salen del stagin se encuentran casi en su totalidad ya aptos para insertarse en el DataMart.

| OTRAS CONSIDERACIONES

No se utilizó la información proveniente de la fuente de datos relacionadas a stock ni a feriados ya que no se solicitan reportes o bien a la empresa no le interesa tener reportes asociados a dichos temas.

ESQUEMA ETL / SSIS



04 | ETL: VERSIONADO

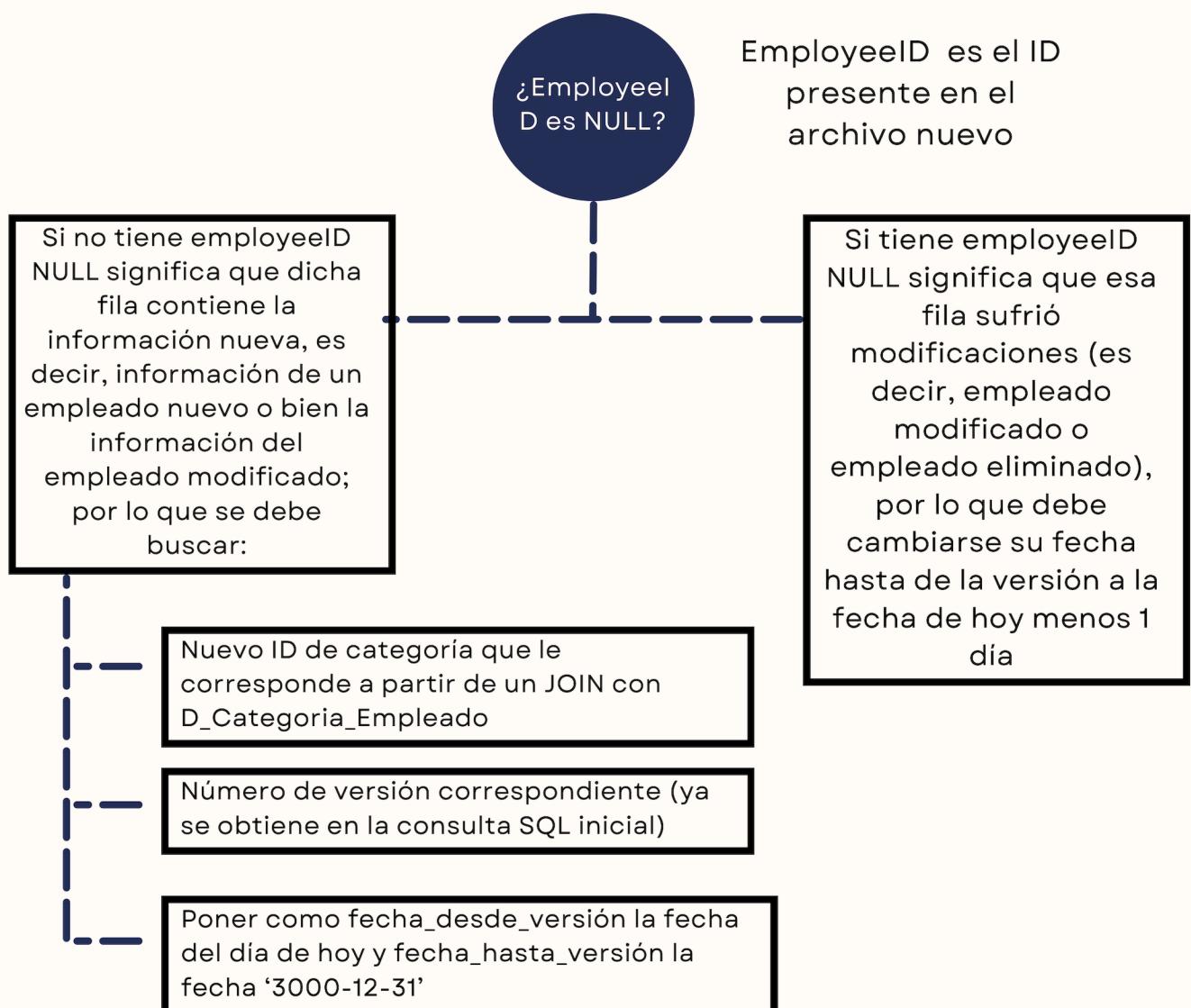


En esta sección se explica la lógica del ETL hecho para el versionado de la D_Empleado a partir de archivos nuevos, donde se detectan cambios: empleados agregados, eliminados o actualizados.

OBTENCIÓN Y COMPARACIÓN

Los datos del nuevo archivo son cargados a un stagi temporal de empleados. Luego se realiza un full join comparando atributo por atributo en coincidencia desde la D_Employees a la tabla Temp_Employees.

De ese join, sólo se seleccionan aquellos que tienen: (legajo empleado null o id empleado null) y (fecha hasta versión = ‘3000-12-31’ o fecha hasta versión null). Ya que el cumplimiento de esas condiciones indica que hubo un cambio sobre la última versión o es una versión de empleado nueva ; luego la lógica de decisión para identificar las acciones es la siguiente:



REPORTES

para los reportes se utilizó la herramienta PowerBI Desktop. Se muestran los reportes requeridos por el departamento de finanzas de la empresa.

*Para ver los reportes, dirigirse al archivo 'Reportes_TDC' de Power BI