



# Universidad Don Bosco

Datawarehouse y Minería de Datos DMD941 G01T

Docente

Ing. Karens Lorena Medrano Mejía

Actividad

## **Desafío Practico 1**

Proyecto Completo ETL: Data Warehouse para Análisis de Ventas Chinook

Alumno

- Jonathan Rafael Señora Reyes SR232918

Fecha de entrega

Domingo 14 de septiembre 2025

## 1. Introducción

En el entorno empresarial contemporáneo, la capacidad de tomar decisiones estratégicas basadas en datos precisos y consolidados es un diferenciador competitivo fundamental. Los sistemas transaccionales en línea (OLTP), como la base de datos "Chinook", están diseñados para la eficiencia operativa del día a día, registrando un alto volumen de transacciones de manera rápida y normalizada. Sin embargo, su estructura no es adecuada para el análisis complejo y la generación de inteligencia de negocio (Business Intelligence), ya que las consultas analíticas sobre estos sistemas pueden ser lentas y complejas.

Para superar esta limitación, el presente proyecto aborda la implementación de un Data Warehouse, una solución arquitectónica estándar para el soporte de decisiones. El objetivo principal es construir un repositorio de datos centralizado y optimizado para el análisis, transformando el modelo relacional normalizado de Chinook en un modelo dimensional en estrella (OLAP). Este nuevo modelo permitirá realizar consultas de alto rendimiento sobre grandes volúmenes de datos históricos.

El alcance del proyecto abarca el ciclo de vida completo de un proceso de Inteligencia de Negocio, consolidando la información de ventas para permitir un análisis ágil del rendimiento desde diversas perspectivas clave: cliente, geografía, producto, artista y género musical.

Para lograr este objetivo, se ha seguido una estricta metodología ETL (Extracción, Transformación y Carga) utilizando un conjunto de herramientas estándar de la industria:

- **SQL Server Management Studio (SSMS):** Utilizado para el diseño, creación, administración y consulta tanto de la base de datos de origen como del Data Warehouse de destino.
- **Visual Studio 2022 con SQL Server Integration Services (SSIS):** Empleado como la plataforma central para la orquestación del flujo de datos, la implementación de la lógica de transformación y la automatización de la carga de datos.

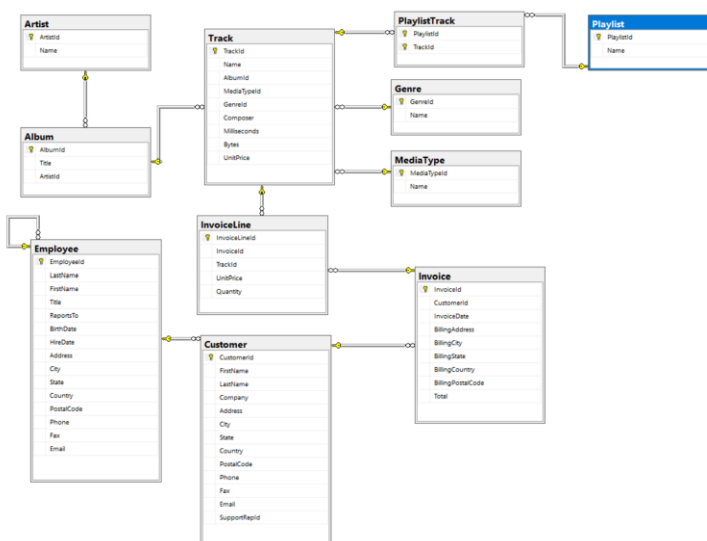
Este documento detalla cada fase del proceso, desde el diseño conceptual del modelo hasta la ejecución de las consultas finales que validan su efectividad.

## 2. Diseño del Modelo Dimensional

Para cumplir con los requisitos analíticos del ejercicio, se optó por un **Modelo Dimensional en Estrella**. Este diseño es el estándar en la industria para proyectos de Business Intelligence y Data Warehousing debido a su simplicidad, alto rendimiento en consultas complejas y facilidad de comprensión para los usuarios finales.

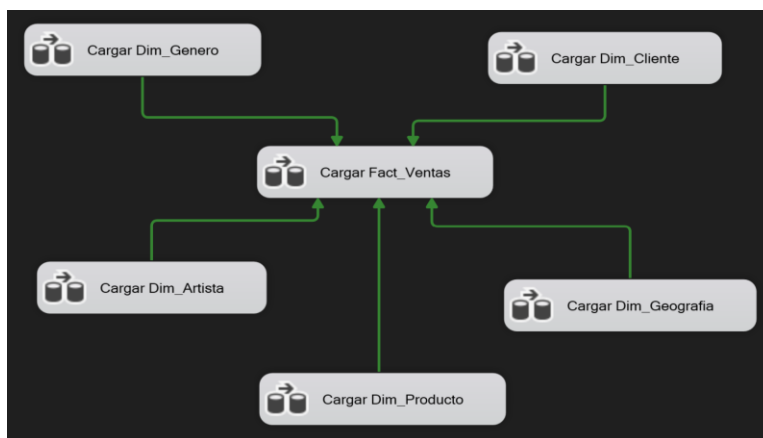
El modelo consta de una tabla de hechos central que almacena las métricas numéricas del negocio y se conecta a múltiples tablas de dimensiones que proveen el contexto descriptivo necesario para el análisis.

**Base de datos normalizada antes de la implementación:**



### 2.1 Diagrama del Modelo en Estrella

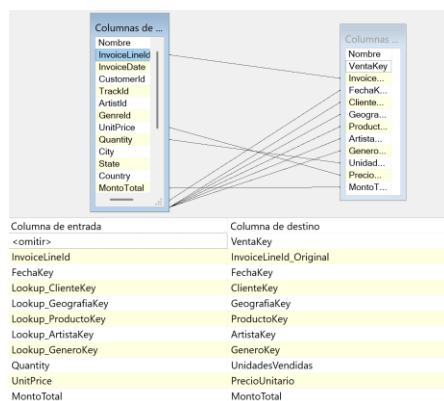
El siguiente diagrama ilustra la relación entre la tabla de hechos y las dimensiones diseñadas para este proyecto:



## 2.2 Descripción de las Tablas

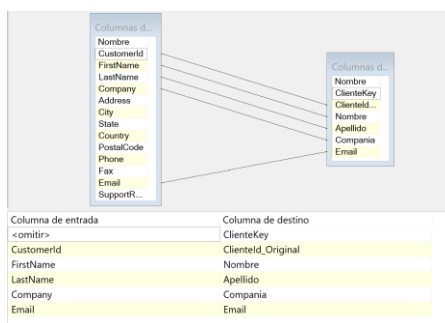
### Tabla de Hechos:

- **Fact\_Ventas:** Es el núcleo de nuestro modelo. Cada fila representa un único evento de negocio: el detalle de una pista musical vendida en una transacción (equivalente a una línea de una factura). Almacena las métricas cuantitativas y las claves foráneas para conectar con las dimensiones.
  - MontoTotal: Medida calculada (Cantidad \* PrecioUnitario).
  - UnidadesVendidas: Medida extraída directamente de la fuente.
  - PrecioUnitario: Medida extraída directamente de la fuente.
  - FechaKey, ClienteKey, GeografiaKey, ProductoKey, ArtistaKey, GeneroKey: Llaves foráneas que conectan el hecho con su contexto descriptivo.

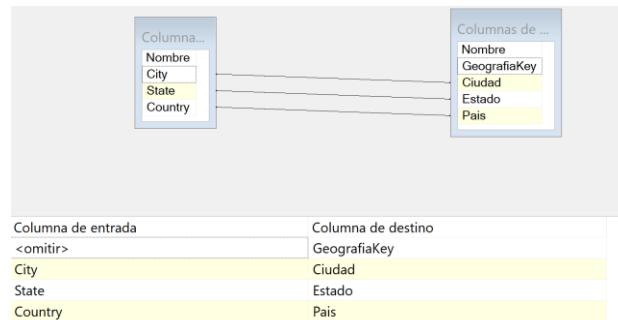


### Tablas de Dimensiones:

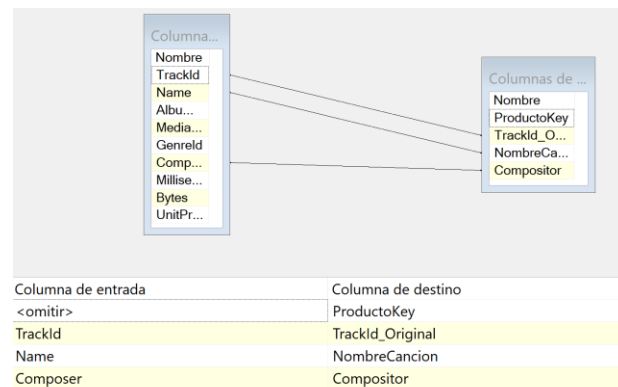
- **Dim\_Cliente:** Describe **quién** realizó la compra. Contiene la información relevante y depurada del cliente.



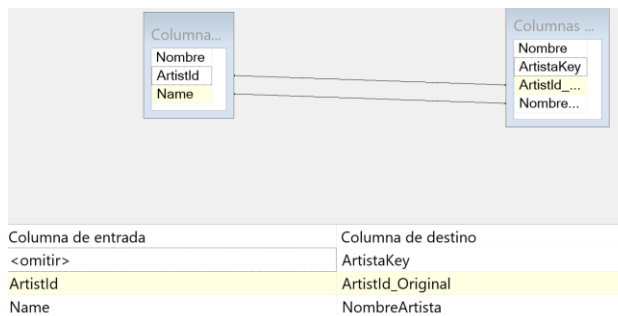
- **Dim\_Geografia:** Describe **dónde** se realizó la compra. Se ha creado a partir de los datos de dirección del cliente para permitir análisis geográficos por ciudad, estado o país.



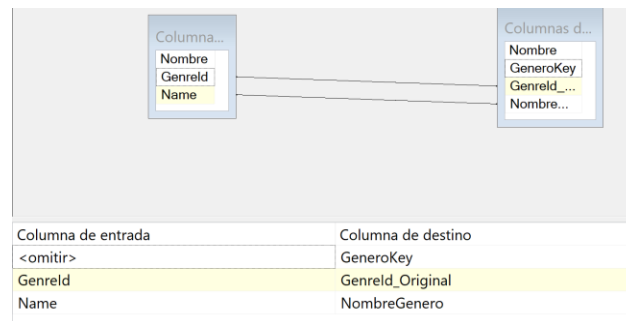
- **Dim\_Producto:** Describe **qué** producto se compró. Contiene la información específica de cada pista musical.



- **Dim\_Artista:** Provee el contexto del artista asociado a cada producto (pista) vendido.

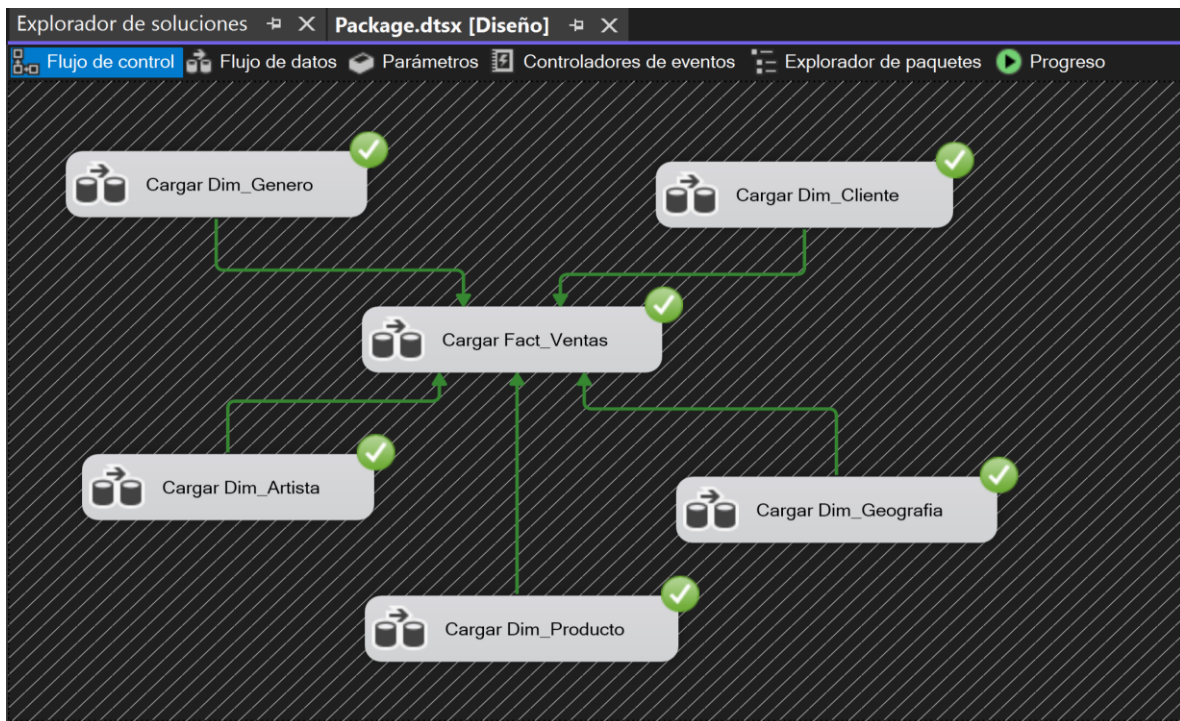


- **Dim\_Genero:** Clasifica cada producto según su género musical, permitiendo análisis de popularidad por estilo.



### 3. Proceso ETL (Extracción, Transformación y Carga)

El proceso ETL se construyó utilizando un proyecto de **SQL Server Integration Services (SSIS)**. El flujo de control fue diseñado para garantizar la integridad referencial, cargando primero todas las dimensiones y, solo después de que todas se completaran con éxito, proceder con la carga de la tabla de hechos.



### 3.1 Extracción

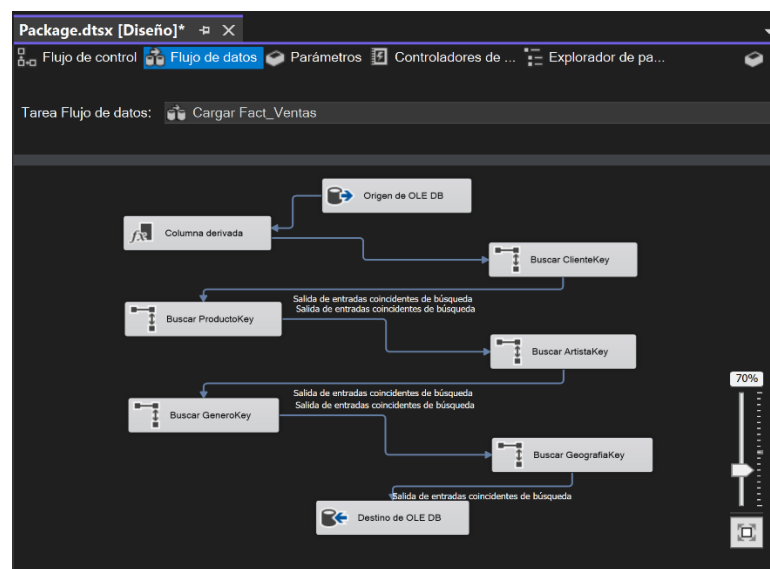
Los datos se extrajeron desde la base de datos Chinook (OLTP) utilizando el componente **Origen de OLE DB**. Para la carga de la tabla de hechos, se consolidó la información de 6 tablas (InvoiceLine, Invoice, Customer, Track, Album, Artist) mediante una única consulta SQL para optimizar el proceso de extracción.

#### Consulta SQL del Origen para Fact\_Ventas:

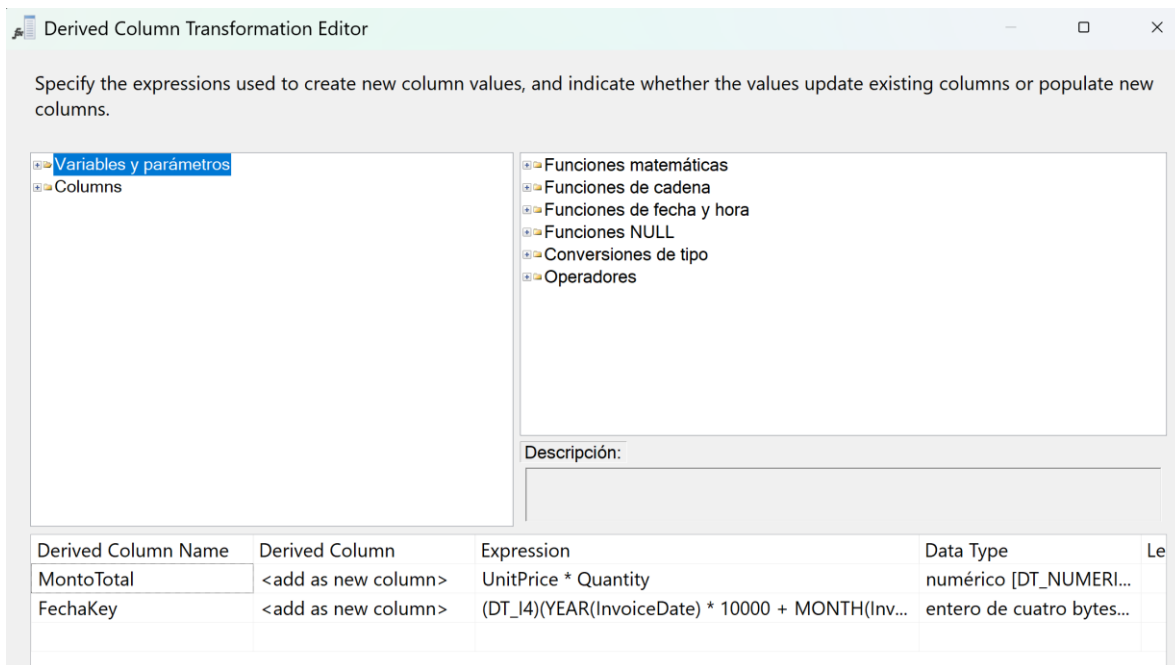
```
SELECT
    il.InvoiceLineId,
    i.InvoiceDate,
    c.CustomerId,
    c.City,
    c.State,
    c.Country,
    t.TrackId,
    al.ArtistId,
    t.GenreId,
    il.UnitPrice,
    il.Quantity
FROM
    dbo.InvoiceLine AS il
INNER JOIN
    dbo.Invoice AS i ON il.InvoiceId = i.InvoiceId
INNER JOIN
    dbo.Customer AS c ON i.CustomerId = c.CustomerId
INNER JOIN
    dbo.Track AS t ON il.TrackId = t.TrackId
INNER JOIN
    dbo.Album AS al ON t.AlbumId = al.AlbumId;
```

### 3.2 Transformación

El flujo de datos para la Fact\_Ventas es el más complejo e incluye las siguientes transformaciones clave para limpiar, enriquecer y preparar los datos:



- **Columna Derivada:** Se utilizó para crear dos nuevas columnas en el flujo de datos:



1. **MontoTotal:** Calculado con la expresión  $\text{UnitPrice} * \text{Quantity}$  para obtener el valor total de cada línea de venta.
  2. **FechaKey:** Creada a partir de InvoiceDate con la expresión  $(\text{DT\_I4})(\text{YEAR}(\text{InvoiceDate}) * 10000 + \text{MONTH}(\text{InvoiceDate}) * 100 + \text{DAY}(\text{InvoiceDate}))$ . Esta conversión a un formato entero YYYYMMDD permite una unión (JOIN) altamente eficiente con la tabla Dim\_Fecha.
- **Búsqueda (Lookup):** Se utilizaron 5 componentes de búsqueda en cadena para enriquecer el flujo. Cada componente toma un identificador de negocio del flujo de entrada (ej. CustomerId) y busca su correspondiente llave subrogada (ej. ClienteKey) en la tabla de dimensión respectiva. Este paso es crucial para traducir los identificadores del sistema de origen a las llaves del Data Warehouse, construyendo así las relaciones del modelo en estrella.



### 3.3 Carga

La carga de datos en las tablas de destino se realizó con el componente **Destino de OLE DB**, conectado a la base de datos Chinook\_DW. En la fase de **Asignaciones (Mappings)**, se verificó que cada columna del flujo de datos se insertara en la columna correcta de la tabla de destino. Se validó la carga mediante el conteo de filas, confirmando que las 2,240 líneas de venta originales fueron cargadas exitosamente en la Fact\_Ventas.

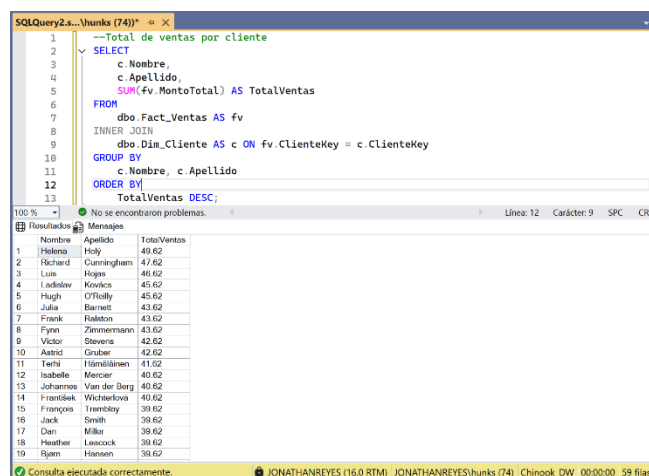
## 4. Consultas de Análisis y Resultados de Negocio

Una vez completado el proceso ETL y poblado el Data Warehouse, se procedió a la fase de explotación de datos. Se ejecutaron las siguientes consultas SQL sobre el modelo dimensional en estrella para validar su eficacia y extraer insights de negocio significativos, tal como lo requiere el ejercicio.

### 4.1.Total de ventas por cliente:

**Objetivo de Negocio:** Identificar a los clientes más valiosos en términos de ingresos generados y obtener un ranking completo del valor de cada cliente.

**Implementación Técnica:** La consulta une la tabla de hechos (Fact\_Ventas) con la dimensión de cliente (Dim\_Cliente) a través de ClienteKey. Agrupa los resultados por nombre y apellido del cliente, sumando la métrica MontoTotal para cada uno. Finalmente, ordena los resultados de forma descendente para destacar a los clientes con mayor contribución.



The screenshot displays a SQL query in the 'SQLQuery2.sql' window and its results in the 'Results' window. The query is titled '--Total de ventas por cliente' and uses an INNER JOIN to combine Fact\_Ventas and Dim\_Cliente tables. The results are ordered by TotalVentas in descending order.

```
1 --Total de ventas por cliente
2 SELECT
3     c.Nombre,
4     c.Apellido,
5     SUM(fv.MontoTotal) AS TotalVentas
6 FROM
7     dbo.Fact_Ventas AS fv
8 INNER JOIN
9     dbo.Dim_Cliente AS c ON fv.ClienteKey = c.ClienteKey
10 GROUP BY
11     c.Nombre, c.Apellido
12 ORDER BY
13     TotalVentas DESC;
```

Nombre	Apellido	TotalVentas
Helena	Poly	49.62
Richard	Cunningham	47.62
Luis	Hogges	46.62
Ladislav	Kovács	45.62
Hugh	O'Reilly	45.62
Julia	Barnett	43.62
Frank	Ralston	43.62
Fynn	Zimmermann	43.62
Victor	Stevens	42.62
Astrid	Gruber	42.62
Terhi	Hämäläinen	41.62
Isabelle	Mercier	40.62
Johannes	Van der Berg	40.62
František	Wichterlová	40.62
Fransjos	Tremblay	39.62
Jack	Smith	39.62
Dan	Miller	38.62
Heather	Leacock	38.62
Ryan	Hansen	38.62

Consulta ejecutada correctamente. JONATHANREYES (16.0 RTM) JONATHANREYES(hunks (74) Chinook\_DW 00:00:00: 59 filas

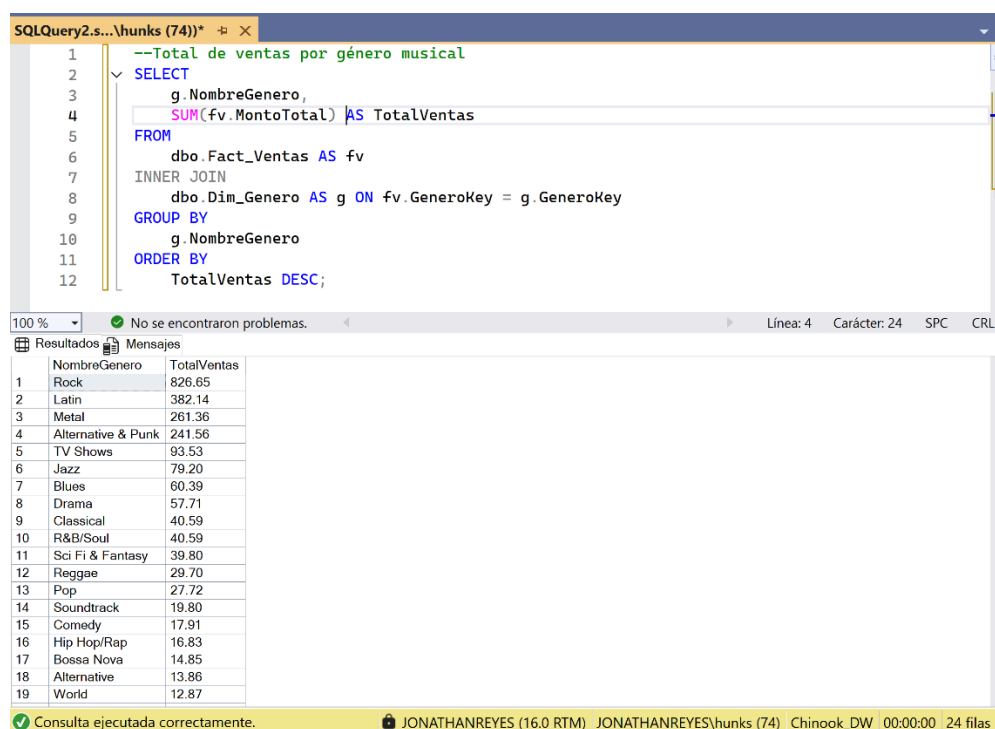
**Análisis del Resultado (59 filas):** La consulta arrojó 59 filas, lo que corresponde al número total de clientes únicos en la base de datos de origen que han realizado al menos una compra. El resultado valida que el ETL ha procesado correctamente la información de todos

los clientes y permite a la empresa identificar rápidamente a sus clientes de mayor valor para futuras campañas de marketing o programas de fidelización.

## 4.2. Total de ventas por género musical:

**Objetivo de Negocio:** Analizar el rendimiento de ventas de los diferentes géneros musicales para entender las preferencias del mercado y tomar decisiones sobre el catálogo de productos.

**Implementación Técnica:** Se realiza una unión entre Fact\_Ventas y Dim\_Genero. La agrupación se hace por NombreGenero, sumando el MontoTotal para obtener una visión consolidada del ingreso por cada categoría musical.



The screenshot displays a SQL query window titled 'SQLQuery2.s... (hunks (74))' with a query that calculates total sales by genre. Below the query, the 'Results' pane shows a table with 24 rows and 2 columns: 'NombreGenero' and 'TotalVentas'. The results are sorted in descending order of total sales. The status bar at the bottom indicates the query was executed successfully and returned 24 rows.

```
1  --Total de ventas por género musical
2  SELECT
3      g.NombreGenero,
4      SUM(fv.MontoTotal) AS TotalVentas
5  FROM
6      dbo.Fact_Ventas AS fv
7  INNER JOIN
8      dbo.Dim_Genero AS g ON fv.GeneroKey = g.GeneroKey
9  GROUP BY
10     g.NombreGenero
11 ORDER BY
12     TotalVentas DESC;
```

NombreGenero	TotalVentas
Rock	826.65
Latin	382.14
Metal	261.36
Alternative & Punk	241.56
TV Shows	93.53
Jazz	79.20
Blues	60.39
Drama	57.71
Classical	40.59
R&B/Soul	40.59
Sci Fi & Fantasy	39.80
Reggae	29.70
Pop	27.72
Soundtrack	19.80
Comedy	17.91
Hip Hop/Rap	16.83
Bossa Nova	14.85
Alternative	13.86
World	12.87

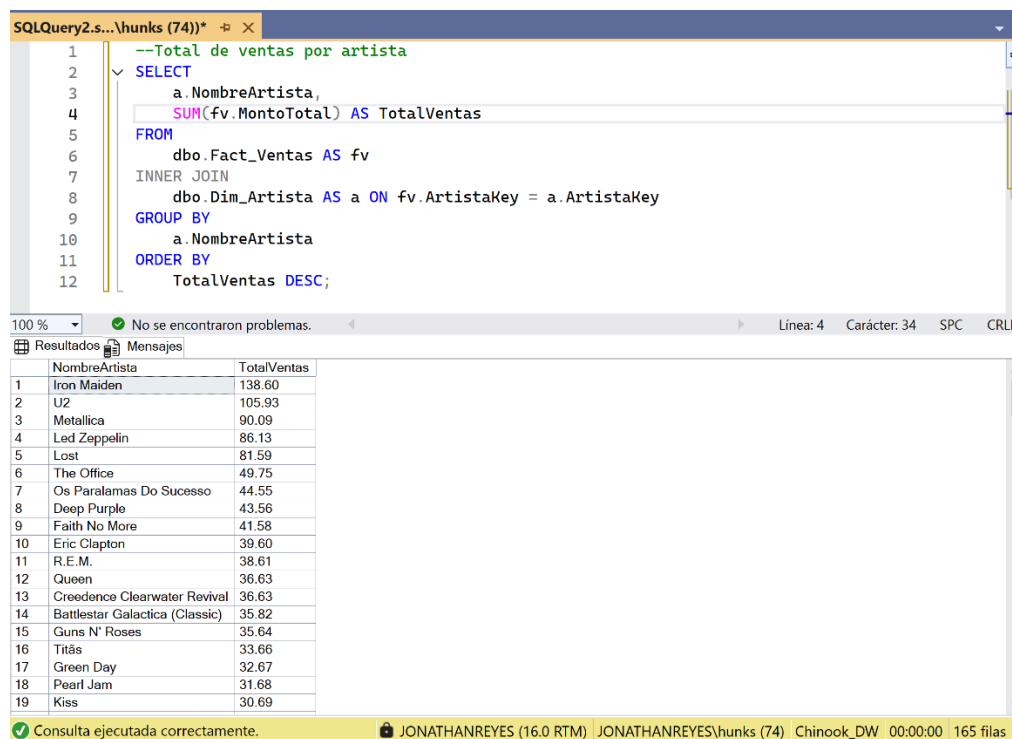
Consulta ejecutada correctamente. JONATHANREYES (16.0 RTM) JONATHANREYES\hunks (74) Chinook\_DW 00:00:00 24 filas

**Análisis del Resultado (24 filas):** El resultado de 24 filas indica que, de los 25 géneros cargados en la dimensión, uno de ellos no ha registrado ninguna venta. Esta es una visión de negocio crucial que no es evidente en el sistema original: permite identificar no solo los géneros más populares (como el Rock, que domina las ventas), sino también aquellos que no están generando ingresos, lo que podría llevar a decisiones estratégicas sobre marketing o inventario.

### 4.3. Total de ventas por artista:

**Objetivo de Negocio:** Determinar qué artistas generan la mayor cantidad de ingresos, permitiendo a la empresa enfocar sus esfuerzos de promoción.

**Implementación Técnica:** La consulta une Fact\_Ventas con Dim\_Artista y agrupa por NombreArtista. La suma del MontoTotal revela el valor comercial de cada artista en el catálogo.



The screenshot displays the SQL Server Enterprise Manager interface. The top pane shows a T-SQL query titled "SQLQuery2.s... \hunks (74))". The query is as follows:

```
1  --Total de ventas por artista
2  SELECT
3      a.NombreArtista,
4      SUM(fv.MontoTotal) AS TotalVentas
5  FROM
6      dbo.Fact_Ventas AS fv
7  INNER JOIN
8      dbo.Dim_Artista AS a ON fv.ArtistaKey = a.ArtistaKey
9  GROUP BY
10     a.NombreArtista
11  ORDER BY
12     TotalVentas DESC;
```

The bottom pane shows the results of the query, titled "Resultados". It displays a table with two columns: "NombreArtista" and "TotalVentas". The table contains 19 rows of data, sorted in descending order of total sales.

	NombreArtista	TotalVentas
1	Iron Maiden	138.60
2	U2	105.93
3	Metallica	90.09
4	Led Zeppelin	86.13
5	Lost	81.59
6	The Office	49.75
7	Os Paralamas Do Sucesso	44.55
8	Deep Purple	43.56
9	Faith No More	41.58
10	Eric Clapton	39.60
11	R.E.M.	38.61
12	Queen	36.63
13	Creedence Clearwater Revival	36.63
14	Battlestar Galactica (Classic)	35.82
15	Guns N' Roses	35.64
16	Titãs	33.66
17	Green Day	32.67
18	Pearl Jam	31.68
19	Kiss	30.69

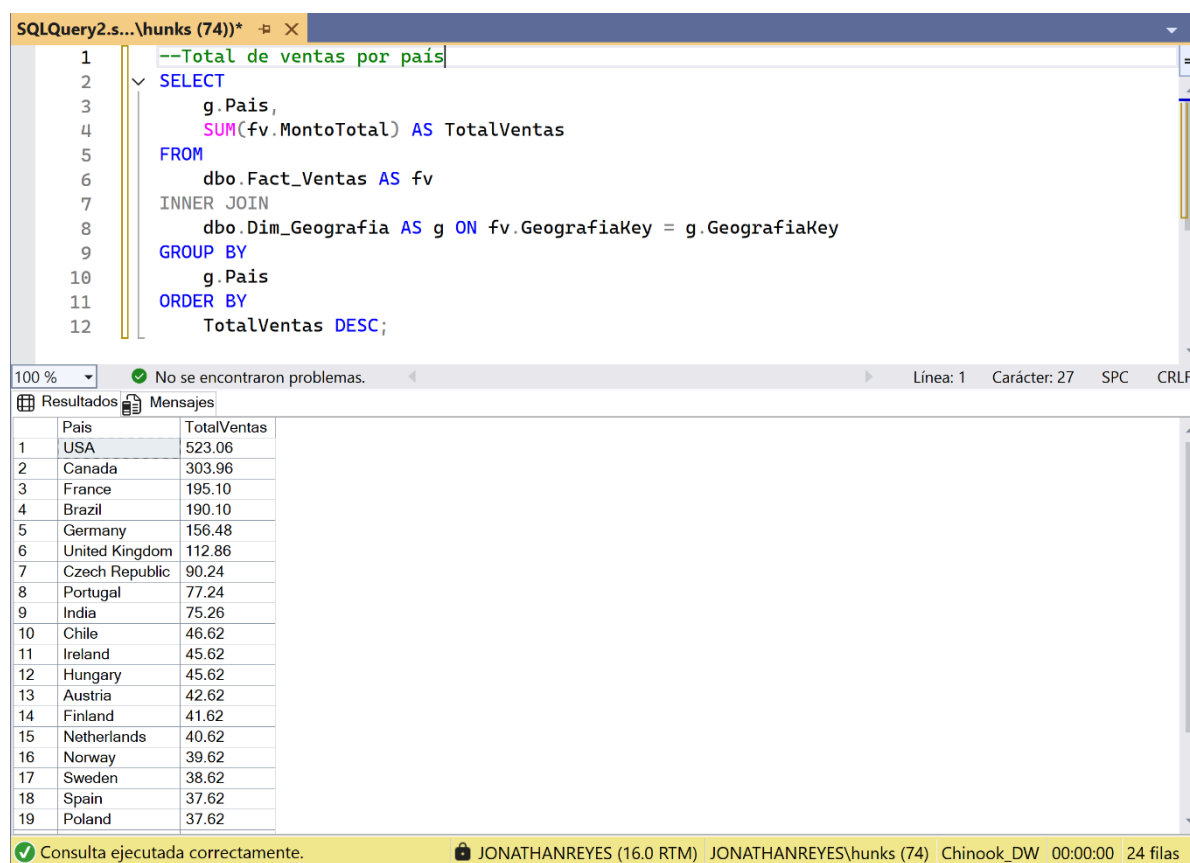
At the bottom of the interface, a status bar indicates: "Consulta ejecutada correctamente. JONATHANREYES (16.0 RTM) JONATHANREYES\hunks (74) Chinook\_DW 00:00:00 165 filas".

**Análisis del Resultado (165 filas):** La consulta revela que, de los 275 artistas presentes en la dimensión, solo 165 han generado ventas. Este es un insight de gran valor (principio de Pareto o regla 80/20), mostrando que una porción del catálogo de artistas es responsable de la totalidad de los ingresos. El ranking descendente permite identificar a los "artistas estrella" de la plataforma.

#### 4.4. Total de ventas por país:

**Objetivo de Negocio:** Evaluar el rendimiento de las ventas a nivel geográfico para identificar los mercados más importantes y las oportunidades de crecimiento.

**Implementación Técnica:** Se une la tabla de hechos Fact\_Ventas con la dimensión Dim\_Geografia. La agrupación por Pais y la suma del MontoTotal consolida los datos para ofrecer una visión macro de la distribución geográfica de los ingresos.



The screenshot displays the SQL Server Enterprise Manager interface. The top pane shows a query titled "SQLQuery2.s... \hunks (74))". The query is as follows:

```
--Total de ventas por país
SELECT
    g.Pais,
    SUM(fv.MontoTotal) AS TotalVentas
FROM
    dbo.Fact_Ventas AS fv
INNER JOIN
    dbo.Dim_Geografia AS g ON fv.GeografiaKey = g.GeografiaKey
GROUP BY
    g.Pais
ORDER BY
    TotalVentas DESC;
```

The bottom pane shows the results of the query, which are displayed in a table with 24 rows. The table has two columns: "Pais" and "TotalVentas". The results are sorted in descending order of "TotalVentas".

	Pais	TotalVentas
1	USA	523.06
2	Canada	303.96
3	France	195.10
4	Brazil	190.10
5	Germany	156.48
6	United Kingdom	112.86
7	Czech Republic	90.24
8	Portugal	77.24
9	India	75.26
10	Chile	46.62
11	Ireland	45.62
12	Hungary	45.62
13	Austria	42.62
14	Finland	41.62
15	Netherlands	40.62
16	Norway	39.62
17	Sweden	38.62
18	Spain	37.62
19	Poland	37.62

The status bar at the bottom indicates that the query was executed successfully ("Consulta ejecutada correctamente."). The status bar also shows the user "JONATHANREYES (16.0 RTM)", the database "JONATHANREYES\hunks (74)", the server "Chinook\_DW", and the execution time "00:00:00". The status bar also shows the number of rows returned, "24 filas".

**Análisis del Resultado (24 filas):** El resultado muestra las ventas totales para cada uno de los 24 países donde residen los clientes. Esta consulta es fundamental para la toma de decisiones a nivel de negocio internacional, permitiendo al departamento de marketing enfocar sus presupuestos en los países de mayor rendimiento o diseñar estrategias para penetrar mercados con ventas más bajas.

## 5. Conclusiones

El proyecto ha culminado exitosamente con la implementación de una solución completa de Business Intelligence, desde la ingesta de datos brutos hasta la capacidad de realizar análisis de negocio complejos. La transición de la base de datos transaccional Chinook a un Data Warehouse dimensional (Chinook\_DW) representa un cambio paradigmático desde un sistema optimizado para operaciones (OLTP) a uno optimizado para análisis (OLAP).

### Áreas Clave:

1. **Implementación de un Modelo Dimensional Robusto:** Se diseñó e implementó con éxito un modelo en estrella, que es la arquitectura fundamental para el análisis de datos de alto rendimiento. Este modelo desnormaliza y organiza los datos en hechos cuantificables y dimensiones contextuales, superando las limitaciones de rendimiento inherentes a las consultas complejas sobre esquemas normalizados.
2. **Desarrollo de un Proceso ETL Automatizado:** Se construyó un paquete de SQL Server Integration Services (SSIS) que automatiza por completo el flujo de datos. Este proceso es modular, mantenible y garantiza la integridad de los datos mediante el uso de restricciones de precedencia, asegurando que las dimensiones se carguen antes que la tabla de hechos.
3. **Transformación de Datos de Valor Añadido:** El ETL no solo movió datos, sino que los transformó para añadir valor. Se realizaron operaciones críticas como:
  - **Limpieza y Estandarización:** Se corrigieron formatos de datos (ej. fechas) para garantizar la compatibilidad y consistencia.
  - **Enriquecimiento de Datos:** Se generaron nuevas métricas (MontoTotal) y atributos (FechaKey) que no existían en el sistema original pero que son cruciales para el análisis.
  - **Integración de Datos:** Se consolidó la información de 6 tablas de origen distintas en una única tabla de hechos, creando una "única fuente de la verdad" para el análisis de ventas.