



TRABAJO DOMICILIARIO: BASES DE DATOS APLICADA I					
FECHA:	27/10/2024				
ALUMNO/A:	Dylan Pointis, Facundo Mercado y Lucas Figueredo				
LEGAJO:	B00068767-T1, B00075055-T1, B00054964-T1	DNI:	45679331, 46215289, 42229273		
CURSO:	2024-090-3-A-M	TURNO:	Mañana		
CARRERA:	Ingeniería en Sistemas Informáticos				
PROFESOR/A:	Martín M. Rivas Sarquis				
MODALIDAD:	Domiciliario				

UNIDADES POR EVALUAR DEL PROGRAMA DE LA MATERIA:

- Unidad 2: Modelando un Data Warehouse.
- Unidad 3: Introducción a minería de datos.

CRITERIOS DE RESOLUCIÓN

Los alumnos/as recibirán la consigna del examen en la fecha de evaluación prevista por el cronograma.

El examen constará de 2 instancias

- 1. Entrega de las consignas y explicación de la metodología de evaluación por parte del docente a los alumnos/as.
 - Duración del examen: 2 (dos) semanas: Los/as alumnos/as deberán entregar la evaluación digital por parte de los alumnos/as al docente.
- 2. Defensa coloquial oral grupal.

INTRODUCCIÓN (100 PTS)

Debe obtener al menos 60/100 para la aprobación de esta parte.

Para llevar adelante esta evaluación domiciliaria grupal, se pensó la siguiente dinámica a desarrollar de 2 (dos) semanas de trabajo. El cual los alumnos tendrán que investigar y llevar adelante la consigna. La evaluación finalizara con una exposición Grupal Oral en defensa y explicación del Trabajo desarrollado.





SAKILA DATABASE

https://github.com/uai-argentina/bda-2024-q2/tree/main/sakila-db

Fue construido originalmente para el motor MySQL, por lo que los nombres de las tablas y los campos están todas en minúscula y cada separación de palabra es dividida por un guion bajo. Esta convención no es la que recomienda Microsoft.

Considérese que su **grupo de trabajo** elegido representa una consultora de sistemas especialista en el desarrollo y construcción de **Data Warehouse** y **Minería de datos**. La cual es contratada para analizar el negocio de **Sakila** que viene funcionando hace más de 10 años. Este simula un sistema de gestión de alquiler de películas y proporciona un entorno que representa las operaciones diarias de una tienda de alquiler de películas.

Los alumnos deberán utilizar como usuario clave del negocio (**Key User**) a <u>ChatGPT</u>, el cual le responderá las consultas funcionales y técnicas del modelo de datos de Sakila. Con el fin de entender y comprender el negocio que se está relevando.

El proyecto se llevará adelante en Fases/Etapas y se debe ir documentando cada una de ellas hasta llegar a la construcción de un **Data Warehouse**.

El resultado final junto con la documentación y todo el proyecto debe ser subido a un **Github** provisto por la universidad o el alumno. De carácter público, que contendrá los siguientes artefactos.

- Documentación del proyecto.
- Solución de Visual Studio.
- Script de creación de Esquemas y objetos necesarios para el Data Warehouse.
- Información sobre el grupo de trabajo, materia, recomendaciones y conclusión.

Lineamientos Generales

- 1. Descarga los Script de SQL de creación del esquema de datos y el de inserción de datos de ejemplo.
- 2. A partir de la base de datos Sakila (OLTP) de datos transaccionales, será necesario relevar el modelo e identificar y construir un Data Warehouse del mismo. Para esto se utilizará la solución de Integration Services (SSIS) para la construcción de ETL.





- a. Utilizar las convenciones recomendadas por Microsoft.
 https://github.com/microsoft/sql-server-samples/tree/master/samples/databases
- Utilizar Visual Studio con la extensión de Integration Services para el desarrollo de ETL.
- c. Utilizar una base de datos de Staging Area para la construcción de los ETL.
- 3. Construir los **Data Marts** que crean convenientes.
- 4. **Documente** todos los pasos y fases de la construcción del Data Warehouse.
- 5. Elegir un Data Mart y construir un CUBO multidimensional con la solución de **Analysis Services**.
- 6. Utilizar **Power BI Desktop** para conectar al Cubo construido en el punto anterior. El mismo se podrá utilizar desde la herramienta.

Links Útiles

- Analysis Service
 - Visual Studio 2019

https://marketplace.visualstudio.com/items?itemName=ProBITools.MicrosoftAnalysisServices ModelingProjects

o Visual Studio 2022

https://marketplace.visualstudio.com/items?itemName=ProBITools.MicrosoftAnalysisServices ModelingProjects2022

- Integration Service
 - o Visual Studio 2019

https://marketplace.visualstudio.com/items?itemName=SSIS.SqlServerIntegrationServicesProjects

o Visual Studio 2022

 $\frac{https://marketplace.visualstudio.com/items?itemName=SSIS.MicrosoftDataToolsIntegrationSer}{vices}$

- SQL Server Data Tools
 - https://learn.microsoft.com/en-us/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-15
- SQL Server Developer Edition
 - o SQL Server 2022 Developer Edition
 - o SQL Server 2019 Developer Edition
- Power BI Desktop
 - o https://www.microsoft.com/en-us/download/details.aspx?id=58494



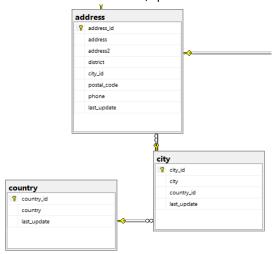


Documentación Data Warehouse

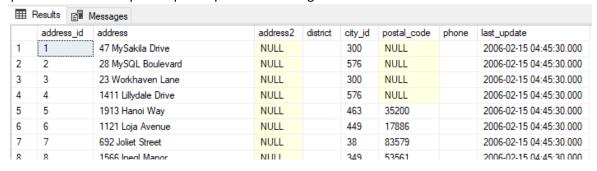
Etapa I: Análisis de los Sistemas de Información

En esta primera etapa, se analizó la fuente de datos para comprender su integración y evaluar su calidad. Es importante encontrar datos incompletos o errores y comprenderlos para garantizar que el Data Warehouse se construya sobre una base sólida de datos confiables y bien integrados.

Durante el análisis de los datos originales pudimos observar que la tabla "address" presenta una clave foránea con ciudad, que a su vez se conecta con país



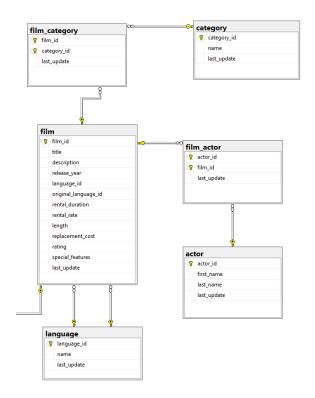
Asimismo, notamos que las columnas "address2", "district" y "phone" no son utilizadas, por lo que son datos incompletos que no pueden ser integrados correctamente en el Data WareHouse



Con respecto a la tabla Film, se analizó cómo se relaciona con otras tablas. Puede tener un lenguaje y un lenguaje original, puede tener una o varias categorías y una o varios actores







Al consultar sus registros, identificamos que el campo "original_language_id" no es utilizado en ningún de ellos, por lo que también se considera como un dato incompleto.



Etapa II: Especificación de Requerimientos

Se contacta con el usuario clave del negocio para entender los requisitos necesarios y relevantes para la organización Sakila. Se identifican los hechos, los datos más importantes para el negocio y la toma de decisiones. Filtrando lo que es realmente útil para los usuarios finales.

En el negocio Sakila las rentas son los datos más importantes, serán los hechos de interés primario a analizar y para la toma de decisiones.

Con respecto a la tabla "film", al estudiarla podemos identificar atributos no relevantes para el negocio como "description", "replacement_cost", "special_features", "last_update" o su relación con actores. Por lo que, al no ser considerados útiles para el negocio, se decidió omitirlos.





En el caso de la tabla "customer", se decidió no utilizar los campos de "store_id", "active", "create_date", "last_update". Consideramos que no aportan valor para el análisis o toma de decisiones.

Para la tabla Address se excluyó el campo "last update".

Para la tabla Staff se decidió no incluir "username", "active", "password", "last_update".

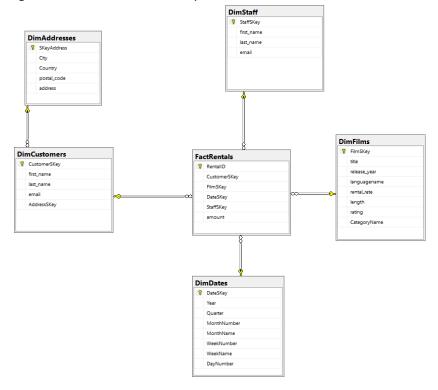
Etapa III: Diseño Conceptual

Se empieza a construir el esqueleto del Data WareHouse, eligiendo la tabla central de hechos "FactRental" e incluyendo su medida "amount".

Se decidió utilizar un esquema copo de nieve, o en inglés snowflake, para obtener un nivel parcial de normalización, una estructura más organizada y evitar cierta duplicación de datos. Se crean jerarquías de dimensiones, conectando una dimensión con otra.

Se crean las dimensiones de "DimCustomers", "DimFilms", "DimDates", "DimAddress" y "DimStaff" para detallar los hechos del negocio.

Se optó por obtener el nombre del lenguaje e incluirlo directamente en la Dimension Films, en lugar de hacer una dimensión aparte.







Etapa IV: Refinamiento y Validación

Se refinó el esquema diseñado en la etapa anterior, para obtener un Data WareHouse que funcione de manera más eficiente y se adapte mejor a las necesidades de la organización.

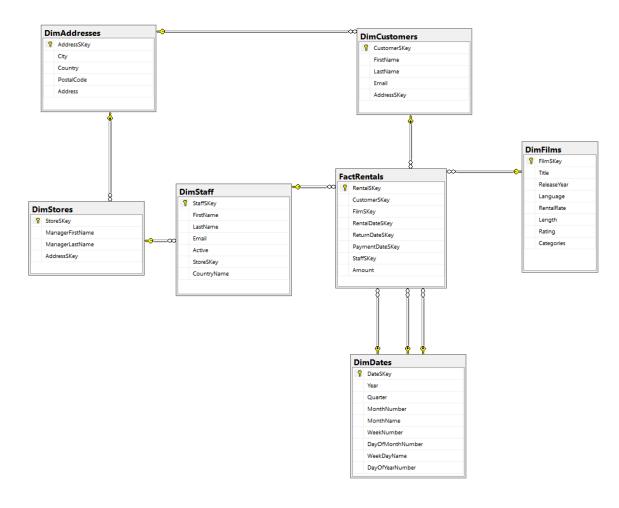
Tras realizar las primeras cargas de trabajo preliminares, se optó por representar las categorías de una película concatenarlas en un campo "Categories".

También se implementaron las convenciones y nomenclaturas recomendadas por Microsoft, así como se validaron los tipos datos para que puedan cargarse correctamente los datos.

Etapa V: Diseño Lógico

Se añade la dimensión Store, considerando que es fundamental a la hora de hacer análisis útiles para el negocio. La dimensión "Staff" se conectó con "Store", de esta forma, podemos acceder a los datos de la tienda desde el empleado que participó en la renta. Así como también, obtener información sobre los empleados de cada tienda.

Además, se optó por agregar un atributo "CountryName" a la dimensión Staff, pues representa el único dato geográfico útil para los empleados de la tienda.







Etapa VI: Diseño Físico

Se implementó el diseño conceptual llevado a cabo en las etapas anteriores en un entorno físico en SQL Server, utilizando el modelo Snowflak, configurando las tablas y sus relaciones en base al diseño definido, consiguiendo cierto nivel de normalización.

Asimismo, se configuraron todos los tipos de datos para que puedan ser completamente compatibles entre la base de datos transaccional y el Data Warehouse construido.

Tras completar la implementación física, se realiza el proceso ETL para poblar las tablas del Data Warehouse con datos históricos, útiles para la toma de decisiones y consultas analíticas en cubos multidimensionales.

PROCESO ETL

Se describe el proceso ETL (Extract, Transform, Load) para cargar los datos desde la base de datos "sakila" al Data WareHouse "DWSakila", realizado en Visual Studio con la extensión de Integration Services.

Antes de cargar los datos en la tabla Fact, primero debemos cargar las dimensiones.

El flujo es el siguiente:

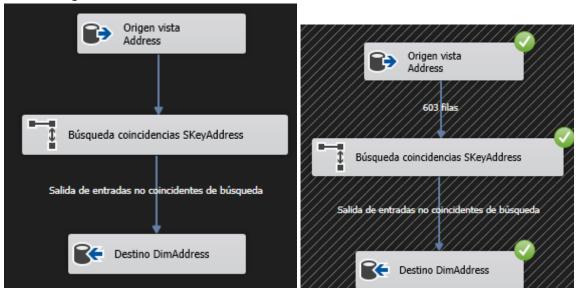






DimAddresses

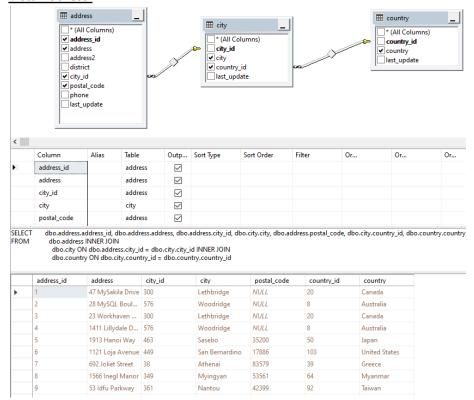
Primero cargamos la dimensión DimAddresses con sus datos



Se implementó una vista para facilitar el acceso a los datos deseados. También se implementó un elemento Búsqueda o "Lookup" para evitar cargar nuevamente registros que ya existan en la dimensión Address del Data WareHouse, evitando conflictos de clave primaria.

Para cargar el resto de las dimensiones, también se utilizaron vistas y el elemento LookUp, con el mismo propósito

Vista Address:

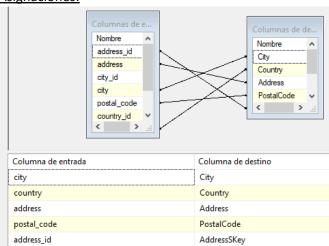




Universidad Abierta Interamericana



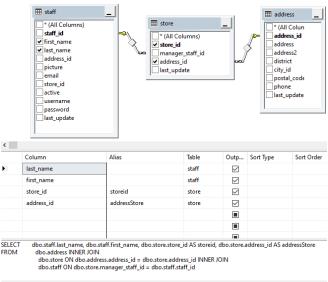
Asignaciones:



DimStore



Vista Store:



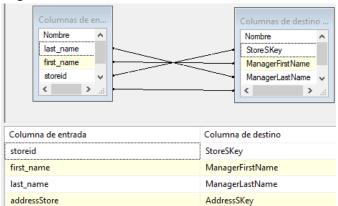
	last_name	first_name	storeid	addressStore
>	Hillyer	Mike	1	1
	Stephens	Jon	2	2



Universidad Abierta Interamericana



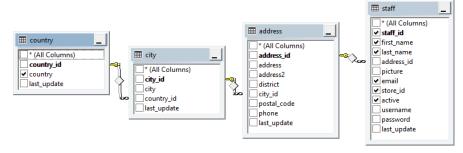
Asignaciones:



DimStaff



Vista Staff:



<									
	Column	Alias	Table	Outp	Sort Type	Sort Order	Filter	Or	Or
)	staff_id		staff	\checkmark					
	first_name		staff	\checkmark					
	last_name		staff	\checkmark					
	email		staff	\checkmark					

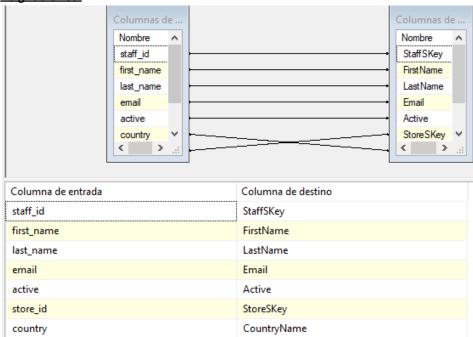
dbo.staff.staff_id, dbo.staff.first_name, dbo.staff.last_name, dbo.staff.email, dbo.staff.active, dbo.country.country, dbo.staff.store_id dbo.address INNER JOIN dbo.staff ON dbo.address_id = dbo.staff.address_id INNER JOIN dbo.city ON dbo.address.city_id = dbo.city.city_id INNER JOIN dbo.country_ON dbo.city.country_id = dbo.country_country_id SELECT

	staff_id	first_name	last_name	email	active	country	store_id
>	1	Mike	Hillyer	Mike.Hillyer@s	True	Canada	1
	2	Jon	Stephens	Jon.Stephens@	True	Australia	2

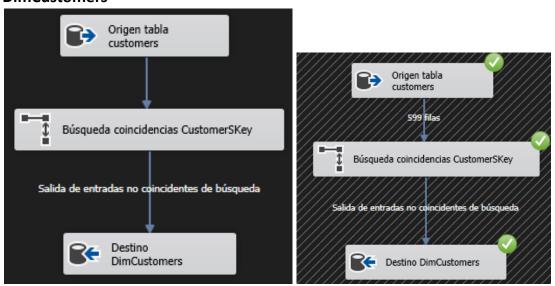




Asignaciones:



DimCustomers

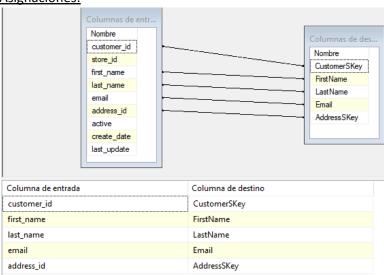


En este caso, no fue necesario utilizar una vista. Directamente utilizamos la tabla para cargar los datos.

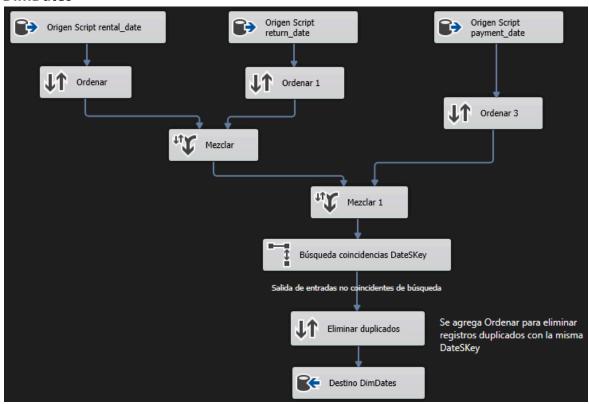




Asignaciones:

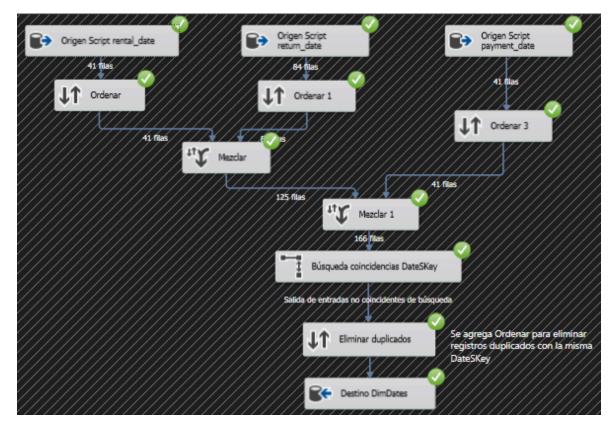


DimDates

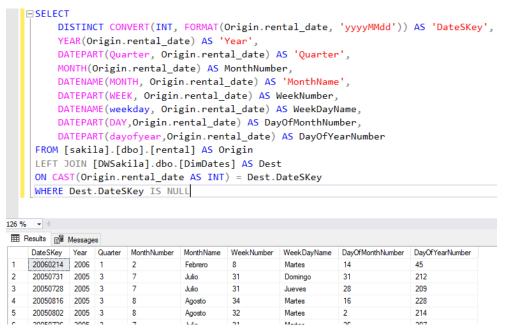








Se utilizó un script para convertir las fechas "rental_date" de tipo DateTime de la base de datos original al tipo de dato INT y así utilizarlas como clave primaria. También para extraer el Año, Cuatrimestre, Número de Mes, Nombre de mes, Número de semana, Nombre del día, Número del día, Número del día del año



Se repitió el mismo script para la fecha de retorno "return_date" y para la fecha de pago

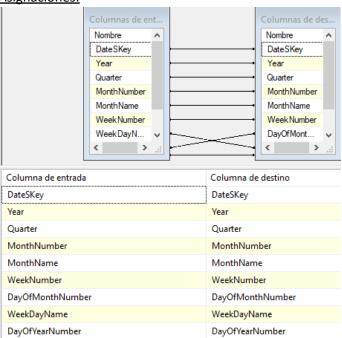




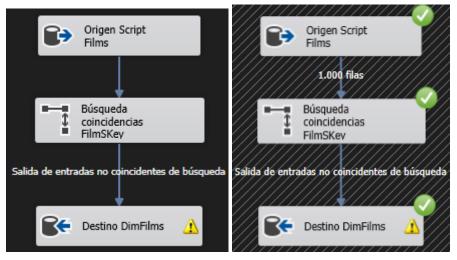
"payment_date".

Se utilizó el elemento mezclar para combinar todos los resultados en un mismo set de datos. El cual luego fue ordenado para eliminar fechas duplicadas.

Asignaciones:



DimFilm



Se utilizó un script para conseguir los detalles de la película, consiguiendo su lenguaje y concatenando sus categorías en el campo "Categories"



Interamericana



Script

```
□ SELECT
       f.film_id,
       f.title,
       f.release_year,

    name AS language,

       f.rental_rate,
       f.length,
       f.rating,
       STRING_AGG(c.name, ', ') AS categories
       FROM film f
       INNER JOIN language 1
       on f.language_id = l.language_id
       INNER JOIN film_category fc ON f.film_id = fc.film_id
       INNER JOIN category c ON fc.category_id = c.category_id
       GROUP BY f.film_id, f.title, f.release_year, l.name, f.rental_rate, f.length, f.rating, c.name
115 % 🕶 🖣
 Results Messages

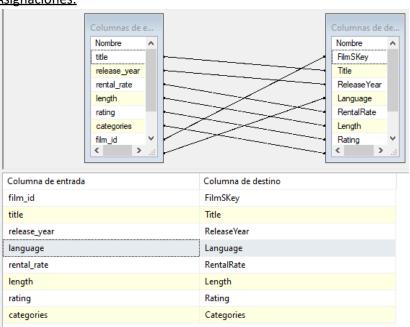
        release_year
        language
        rental_rate
        length
        rating

        2006
        English
        0.99
        86
        PG

        2006
        English
        4.99
        48
        G

      film_id title
                                                                                                 categories
             ACADEMY DINOSAUR
      1
                                                                                                  Documentary
               ACE GOLDFINGER
                                                        English 4.99 48 G Formal Commentary
English 2.99 50 NC-17 Documentary
English 2.99 117 G Homor
English 2.99 130 G Family
 2
                                                                                                 Horror
     3
               ADAPTATION HOLES
 3
                                           2006
               AFFAIR PREJUDICE
 4
      4
                                            2006
               AFRICAN EGG
 5
                                           2006
```

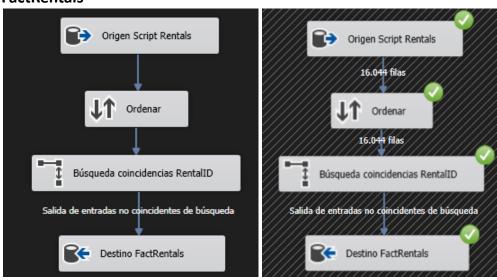
Asignaciones:







FactRentals



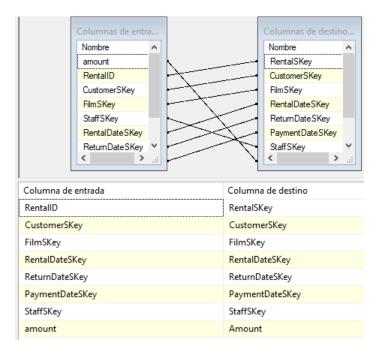
Se utilizó un script para convertir la fecha de renta, de retorno y de pago en INT y así cargarlos en RentalDateSKey, ReturnDateSKey y PaymentDateSKey. Además, se usó para unir con Inventory y conseguir el Film key de la renta

Script □ SELECT dbo.rental.rental id AS RentalID, dbo.rental.customer id AS CustomerSKey, dbo.inventory.film_id AS FilmSKey, dbo.rental.staff_id AS StaffSKey, CONVERT(INT, FORMAT(dbo.rental.rental_date, 'yyyyMMdd')) AS RentalDateSKey, CONVERT(INT, FORMAT(dbo.rental.return_date, 'yyyyMMdd')) AS ReturnDateSKey, CONVERT(INT, FORMAT(dbo.payment.payment_date, 'yyyyMMdd')) AS PaymentDateSKey, dbo.payment.amount FROM dbo.payment INNER JOIN dbo.rental ON dbo.payment.rental_id = dbo.rental.rental_id INNER JOIN dbo.inventory ON dbo.rental.inventory_id = dbo.inventory.inventory_id order by RentalID 115 % Results Messages RentalID CustomerSKey FilmSKey StaffSKey RentalDateSKey RetumDateSKey Payment Date SKey amount 130 80 1 20050524 20050526 20050524 2.99 1 2 459 333 1 20050524 20050528 20050524 2.99 408 3.99 3 3 373 1 20050524 20050601 20050524 4 4 333 535 2 20050524 20050603 20050524 4.99 5 5 222 450 20050524 20050602 20050524 6.99

Asignaciones

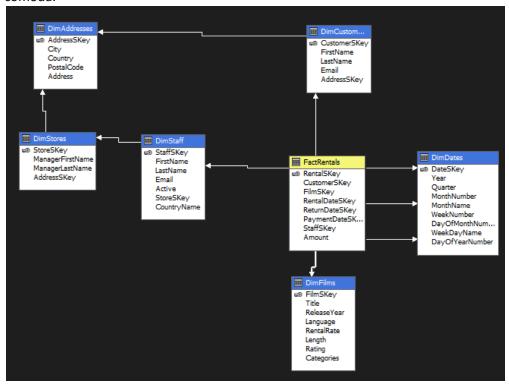






CONSTRUCCION DE CUBO MULTIDIMENSIONAL

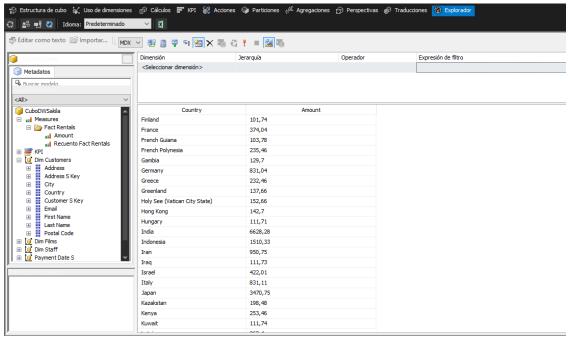
Utilizando la herramienta de Analysis Services se llevó a cabo la construcción del cubo multidimensional, para consultar los datos desde diferentes perspectivas, de una manera cómoda.



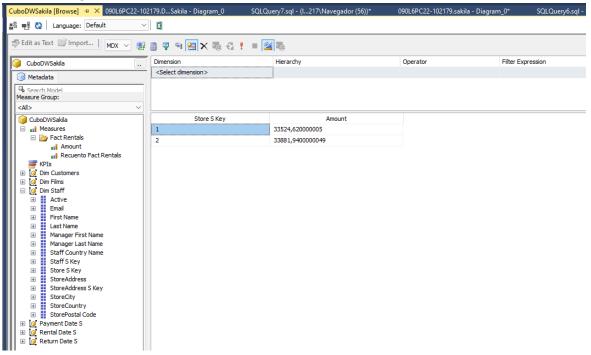








En SQL Management Studio:



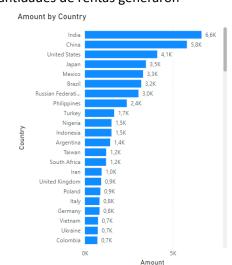




POWER BI

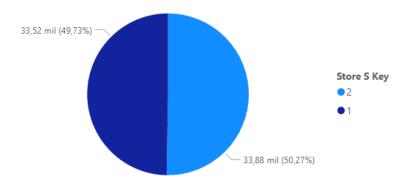
Ejemplos de gráficos posibles a implementar: Paises de clientes que mas ingresos y cantidades de rentas generaron

Country	Amount	Recuento Fact Rentals
India	6.628,28	1572
China	5.798,74	1426
United States	4.110,32	968
Japan	3.470,75	825
Mexico	3.307,04	796
Brazil	3.200,52	748
Russian Federation	3.045,87	713
Philippines	2.381,32	568
Turkey	1.662,12	388
Nigeria	1.511,48	352
Indonesia	1.510,33	367
Argentina	1.434,48	352
Taiwan	1.209,95	305
South Africa	1.204,15	285
Iran	950,75	225
United Kingdom	922,81	219
Poland	877,97	203
Italy	831,11	189
Germany	831,04	196
Vietnam	746,28	172
Ukraine	730,42	158
Colombia	709,41	159
Egypt	694,39	161
Venezuela	683,30	170
Total	67.406,56	16044



Recaudación por tienda

Amount por Store S Key



Recaudación por tienda, año y trimestre

Store S Key	Amount	Year	Quarter
1	7.172,79	2005	2
1	26.133,66	2005	3
1	218,17	2006	1
2	7.280,54	2005	2
2	26.305,39	2005	3
2	296,01	2006	1

Total 67.406,56