## ESCUELA POLITÉCNICA NACIONAL

ESCUELA DE FORMACIÓN DE TECNÓLOGOS

### ANÁLISIS DE DATOS TSDS

ASIGNATURA: ANÁLISIS DE DATOS

PROFESOR:

FECHA:

Ing. Lorena Chulde / Ing. Juan Pablo Zaldumbide

13 – 08 - 2024

PERÍODO ACADÉMICO: 2024-A

## PPROYECTO FINAL – BIMESTRE 2



**Integrantes:**

**Richard Robalino, Angelo Loor, Emilio Erazo**

# 2024-A

**Definición del caso de estudio.**

El proyecto consiste en la integración de diversas bases de datos, tanto relacionales como NoSQL, para crear una infraestructura robusta que permita la consolidación de datos de múltiples fuentes. Se utilizarán PostgreSQL, MongoDB, Redis, y SQL Server como las principales bases de datos, y se establecerán conexiones entre estas utilizando MySQL como intermediario cuando sea necesario. La data consolidada se almacenará en un repositorio en SQL Server para su posterior análisis y visualización a través de Power BI, permitiendo la generación de dashboards y consultas sobre diferentes temas.

**Objetivos General y específicos.**

**Objetivo General**

Desarrollar una infraestructura de base de datos integrada que combine tecnologías SQL y NoSQL, permitiendo la consolidación de datasets en un repositorio central en SQL Server, para su posterior análisis y visualización mediante Power BI.

**Objetivos específicos**

1. Integrar múltiples bases de datos (PostgreSQL, MongoDB, Redis, SQL Server) en un entorno cohesionado.

2. Establecer conexiones eficientes entre las bases de datos SQL y NoSQL utilizando MySQL como intermediario cuando sea necesario.

3. Desarrollar un repositorio central en SQL Server que almacene los datasets consolidados de todas las fuentes.

4. Conectar el repositorio central con Power BI para la creación de dashboards y consultas temáticas.

5. Optimizar el flujo de datos entre las distintas bases de datos para asegurar la integridad y eficiencia en la transferencia de información.

**Descripción del equipo de trabajo y actividades realizadas por cada uno.**

Richard: Encargado de la integración y configuración de las bases de datos relacionales (PostgreSQL y SQL Server). Richard también es responsable de la configuración del repositorio central en SQL Server y su conexión con Power BI.

Angelo: Responsable de la configuración e integración de las bases de datos NoSQL (MongoDB y Redis). Angelo trabaja en la optimización del flujo de datos entre las bases de datos NoSQL y su transferencia al repositorio en SQL Server.

Emilio: Especialista en la intermediación de datos utilizando MySQL. Emilio trabaja en el establecimiento de las conexiones necesarias entre las bases de datos SQL y NoSQL, asegurando que la información fluya correctamente hacia el repositorio central.

* Cronograma de actividades incluído el diagrama de gantt (EXCEL)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ACTIVIDAD** | **Richard Robalino** | **Angelo Loor** | **Emilio Erazo** | **Semana 1** | **Semana 2** | **Semana 3** | **Semana 4** |
| Configuración de PostgreSQL y SQL Server | 2H | 3H |  | X |  |  |  |
| Configuración de MongoDB y Redis |  | 2H | 2H |  | X |  |  |
| Establecimiento de conexiones entre bases de datos |  | 2H | 1H |  |  | X |  |
| Desarrollo del repositorio central en SQL Server | 2H | 3H |  |  |  |  | X |
| Optimización del flujo de datos |  | 2H | 2H |  |  |  | X |
| Pruebas de integración y transferencia de datos |  | 2H | 2H |  |  |  | X |
| Conexión con Power BI y pruebas de visualización | 2H | 3H |  |  |  | X |  |
| |  | | --- | | Revisión final y ajustes |      |  | | --- | |  | |  | 2H | 1H |  |  |  | X |

**Recurso y herramientas utilizadas.**

Power BII

Redis Data Studio

Postgresql

MongoDB

Jupyter Notebook

Sql Server

Studio 3T

Excel

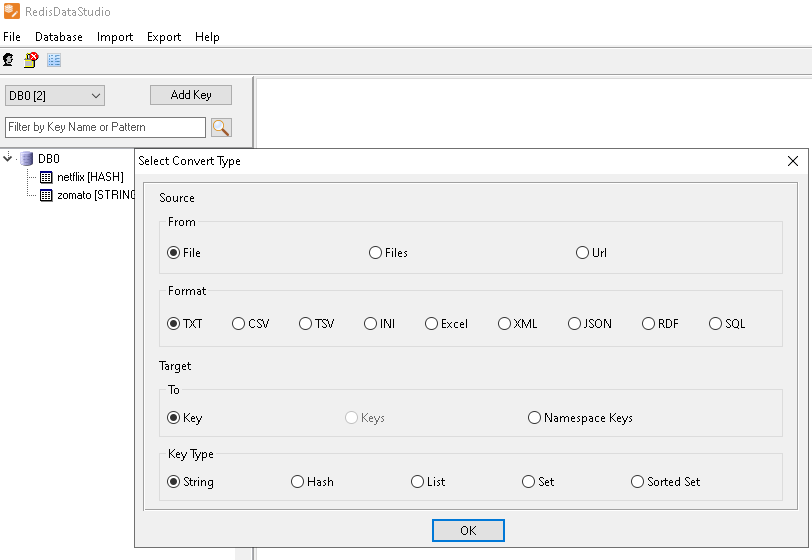
**Arquitectura de la solución.**

Primero abrimos el servidor de redis con en cmd con lo siguiente comandos





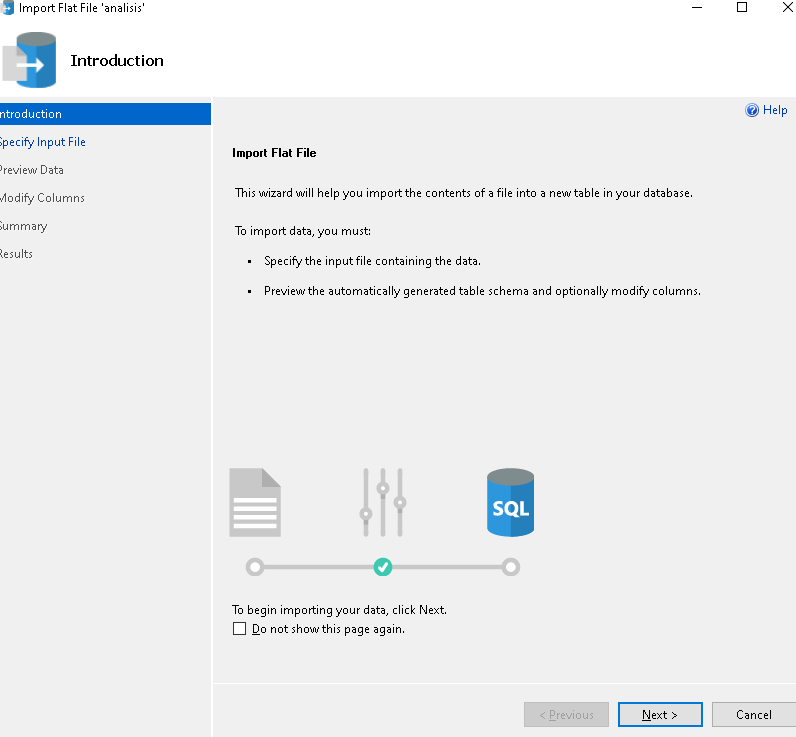
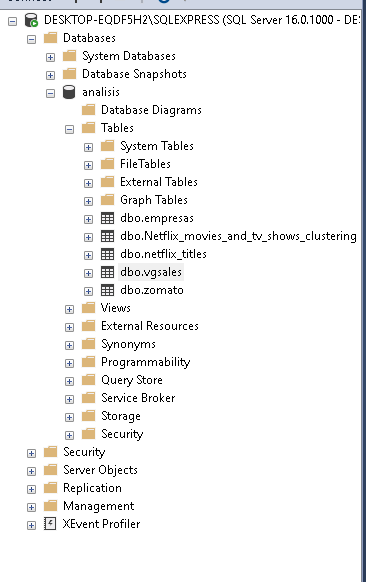
Después podemos insertar datos mediante el programa Redis DataStudio



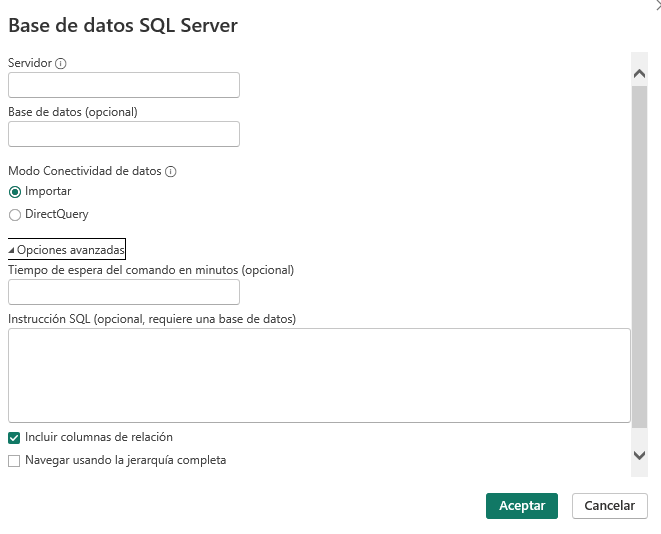
Después conectamos conectar al mongodb desde redis mediante el codigo que se ejecuta en jupyter notebook

Al momento de cargar los datos a mongodb podemos importar los datos en csc y json

Después , esos datos cargamos los datos a sqlserve



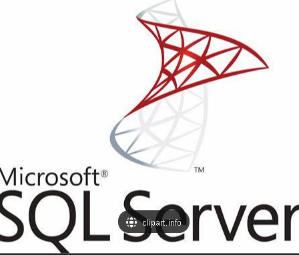
Finamente conectamos los datos que se encuentran en mi sqlserve a powerbi



Ingresamos los datos a mongoDB y descargamos el studio 3T para poder conectarnos sql server y poder migrar los datos creando así un documento de sql.

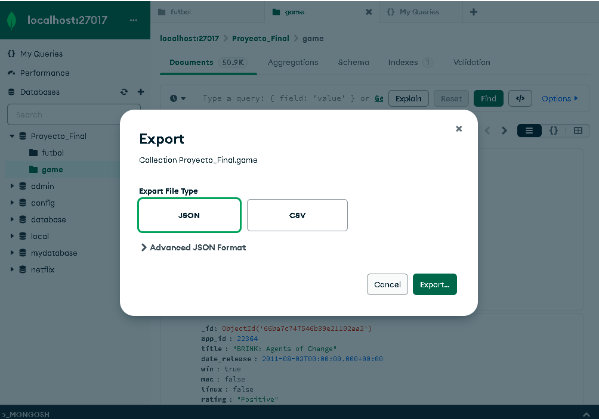
 

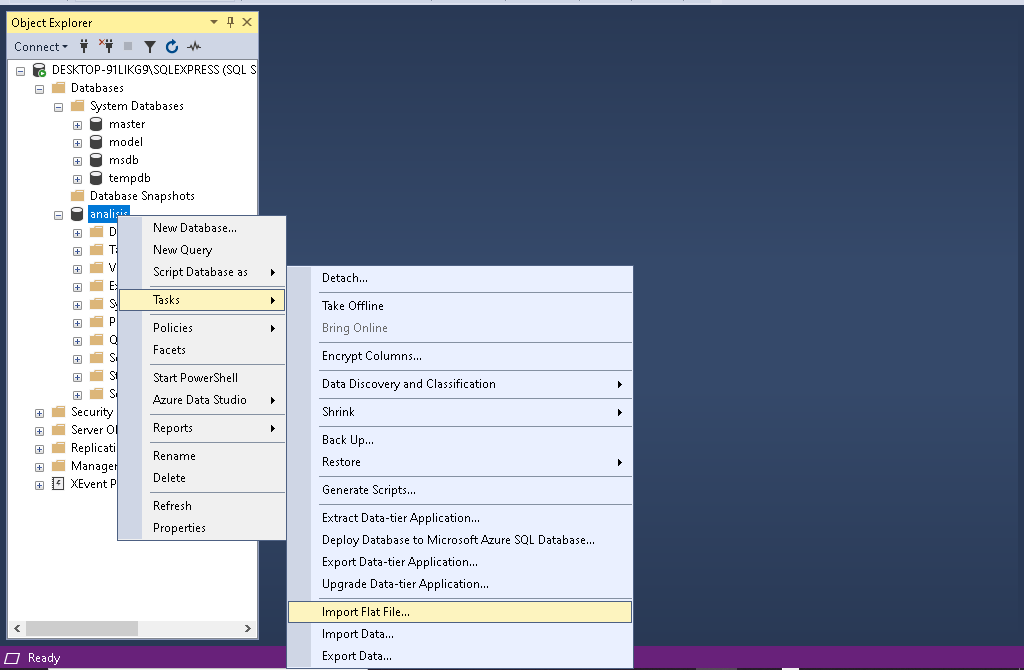
Estructura Principal



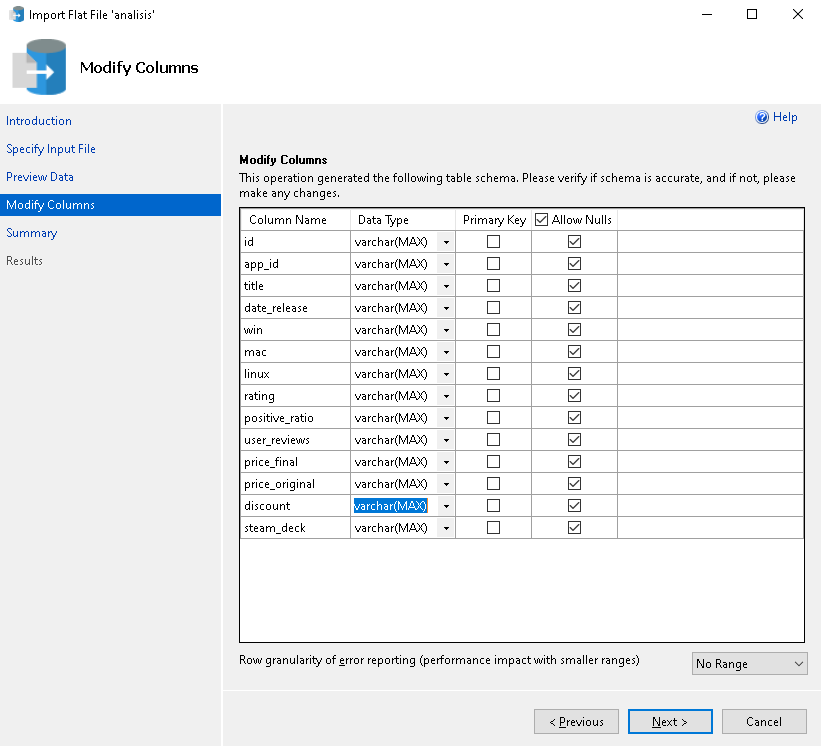
* Extracción de datos.

Importar los datos a mongo db o redis para poder hacer las exportaciones de json o csv.

* 



Importacion de datos en csv o json



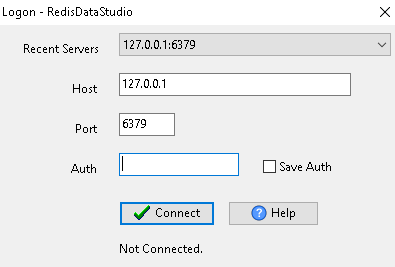
Selección del tipo de dato y los nulls si existen

Redis

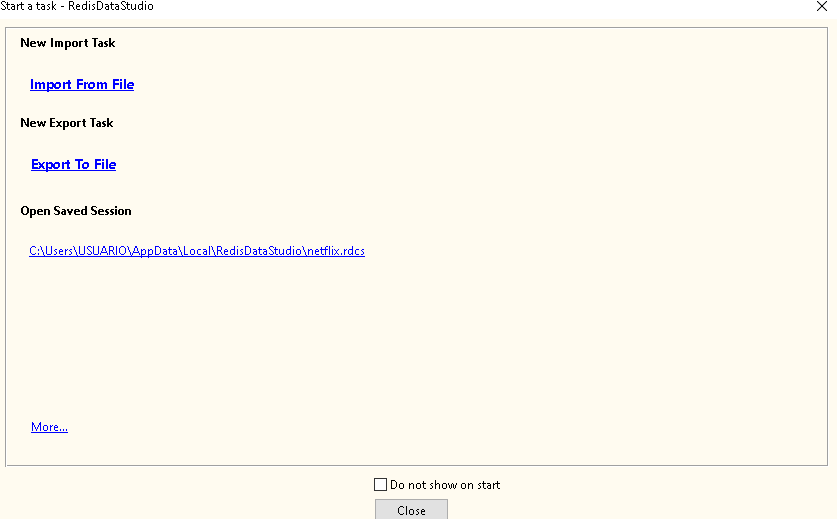
**Importación de datos**

Tenemos que descargar el programa Redis DataStudio para realiza la importación y exportación de datos

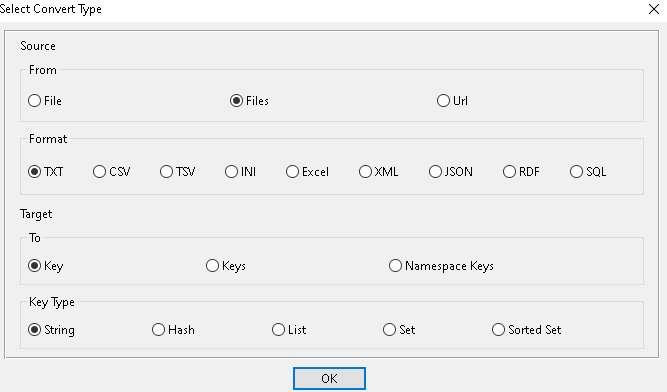
Primero debemos conectarnos a Redis



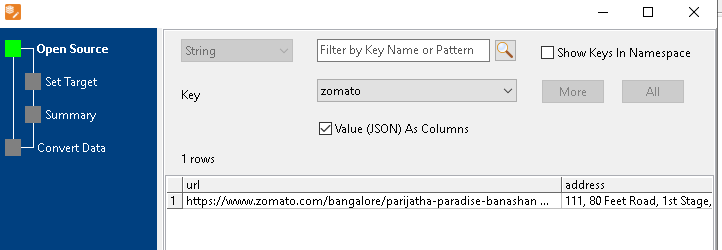
2. Debemos seleccionar importa o exportar

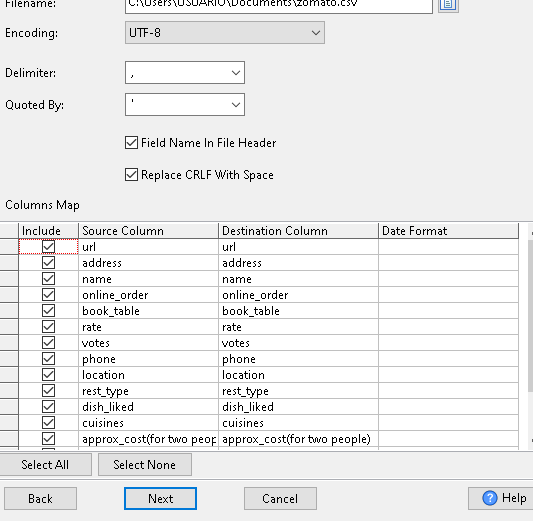


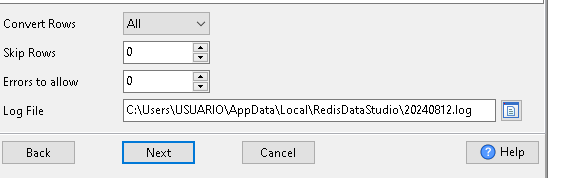
Exportar que podemos escoger en que formato del documento de deseemos subir al redis

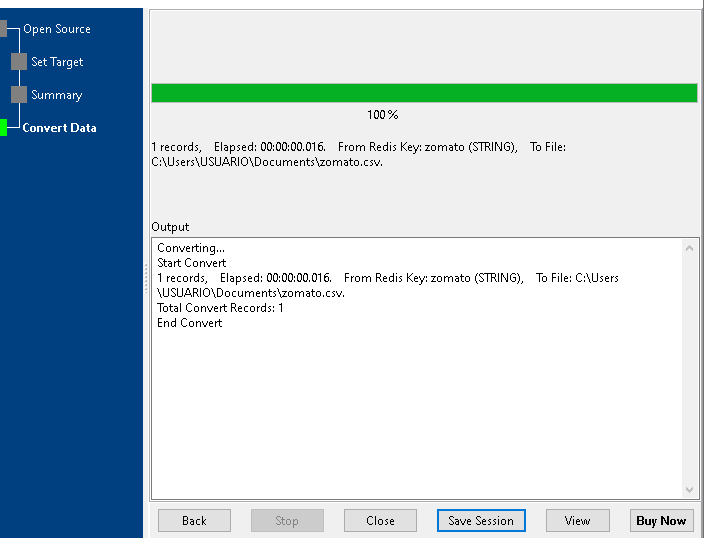


Selección la base que vamos a exportar







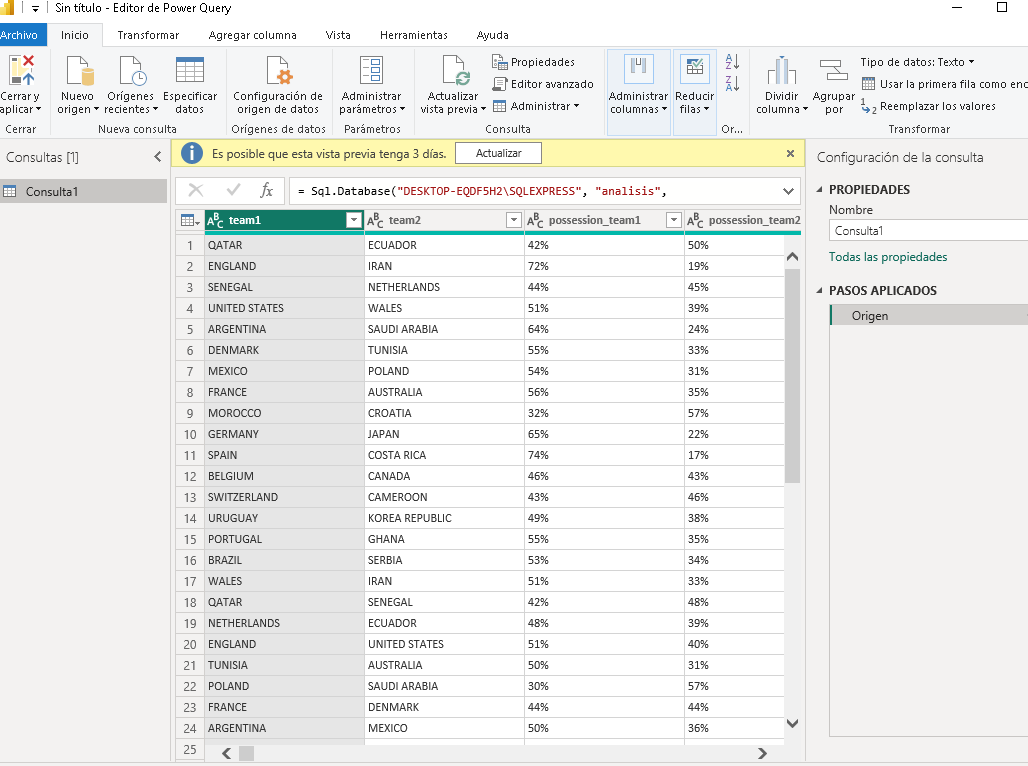


Dada las exportaciones podemos ingresar esos datos a sql server.

**Análisis de información.**

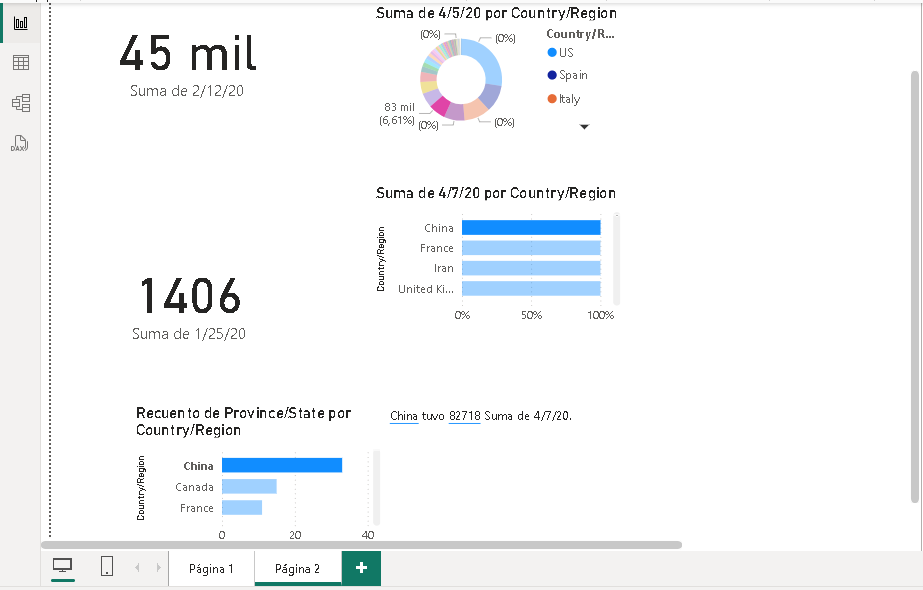
1. Empresa de Ventas de tecnología pose datos 2015 sobre empres que vende producto como computadoras, impresora, laptops, etc.
2. Ventas de videojuegos la data set contiene datos venta de videojuegos por el mayor número de juegos vendidos por nombre o por categorías
3. Series de televisión un data set que contiene la duración por season para temporada
4. Futbol: este data set es sobre la world cup en la cual indica los golos del team 1 o 2 se muestren los tiros de esquina, si ganaron por penaltis y el ganador de la world cup.
5. Game: se muestran los juegas más rankeados, la longitud del game el win del juego más proporcionados.
6. Covid: Noticia de la enfermedad que altero a todo el mundo por días a años de cómo se estuvo propagando la enfermedad, el nivel de contagio y el país más dañado
7. Restaurante: muestra una database que como se lo consultes te va a dar lo que es la localidad y Lo famoso que son
8. Electricidad acorde a los vehículos en WA-USA: Se puede visualizar mapeado y la marca de vehículo que ocupa electricidad en el estado de Washington.
9. Juego de Loteria: Se puede visualizar el conteo de una serie de personas que ganaron
10. Baloncesto: Muestra el salario, posición, puntos anotados de 2p y de 3p puntos de una gran lista de jugadores de la NBA

Limpieza de dato ya que al momento de carga los datos sqlserver le estamos cargado todo con varchar, debemos transforma los datos a sus datos originales como las fecha, datos números, horas, minutos esta opción de transformar datos el mismos power bi no da la función

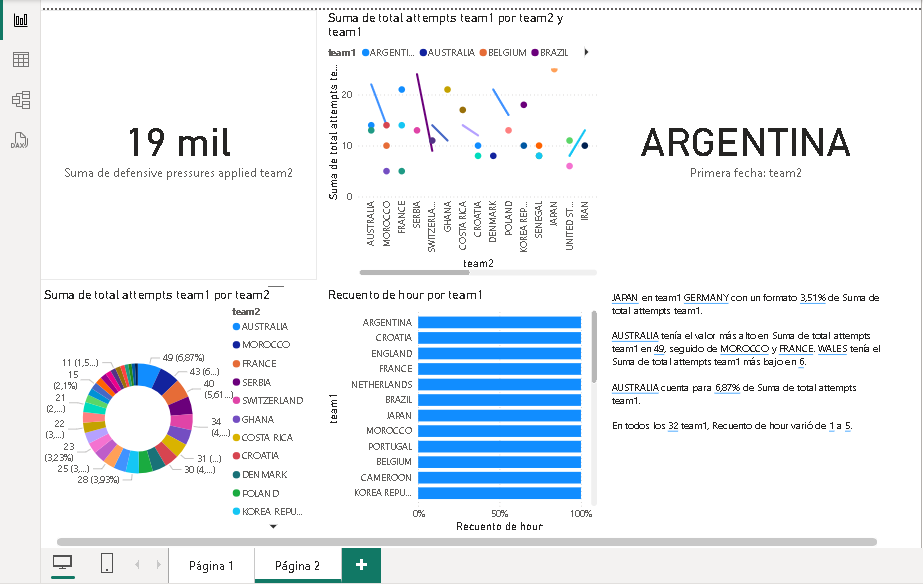


* Visualización de información.

Dashboards de la data set Covid



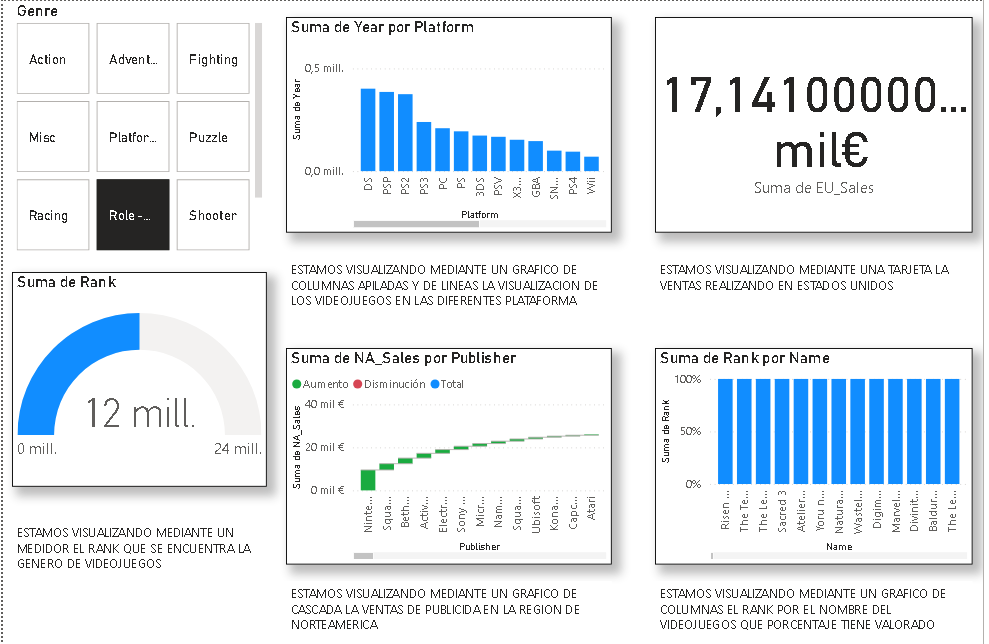
Dashboards de la data set Futbol

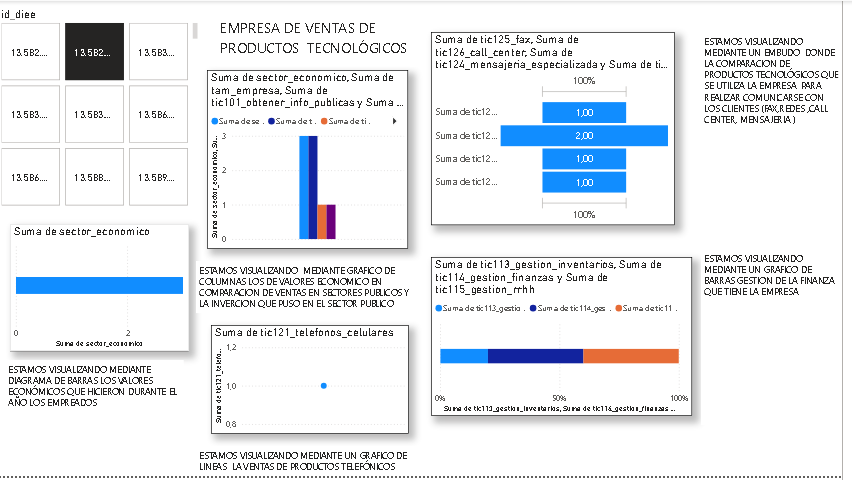


Dashboards de la data set Game

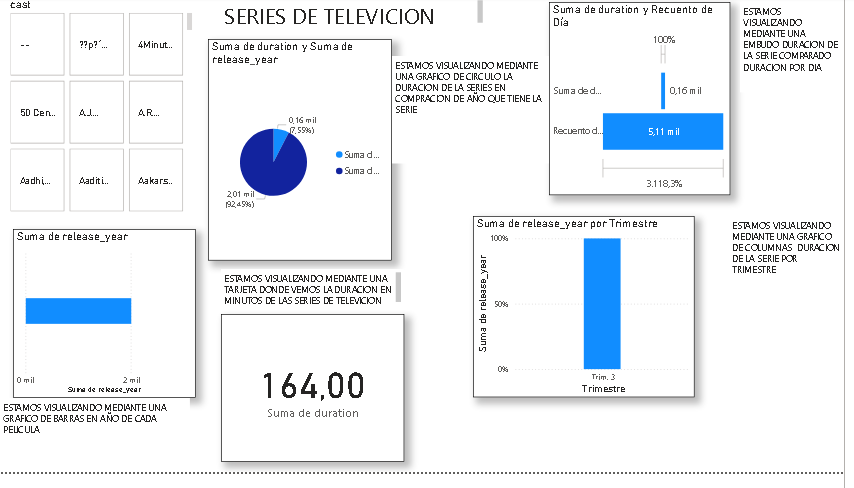


Ventas de Videojuegos

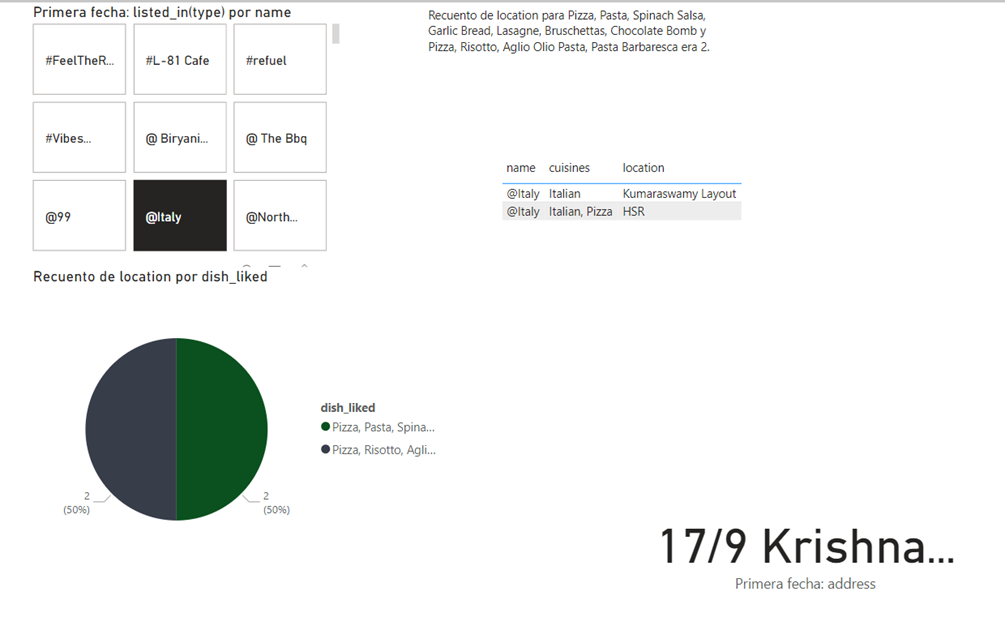
Empresa de ventas de tecnología



Series de Televisión

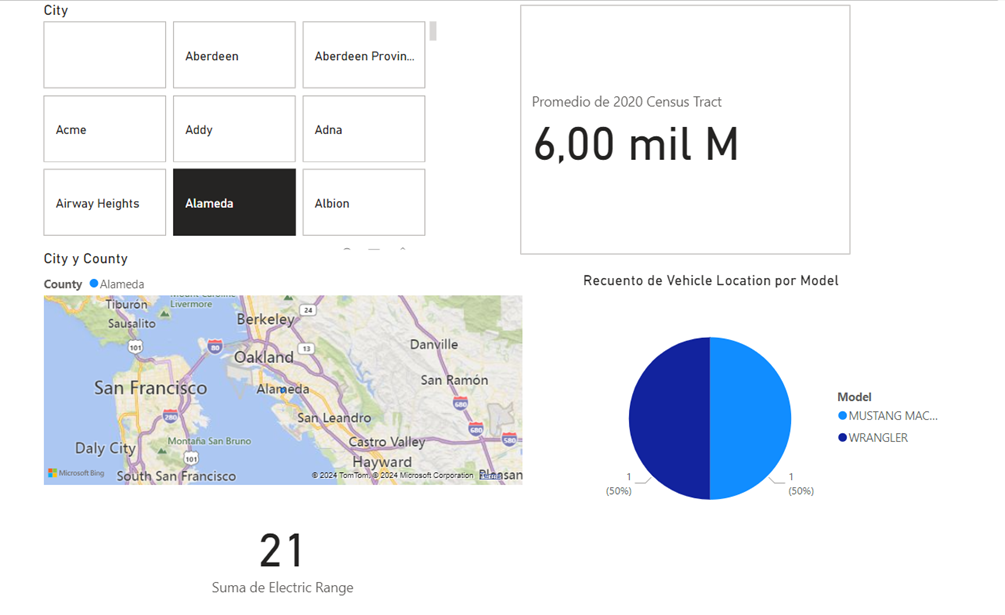


Restaurante



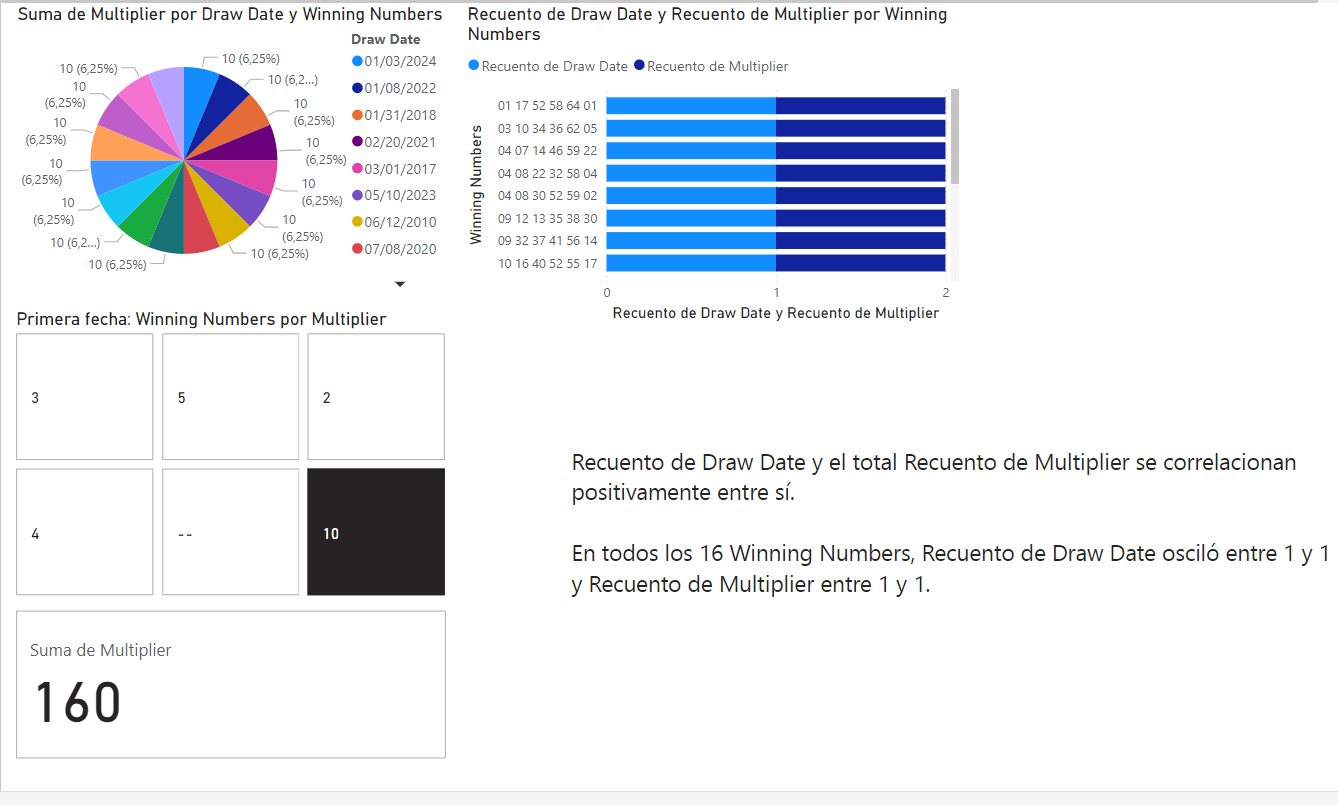
Electricidad acorde a los vehículos en WA-USA

Se puede visualizar el lugar mapeado y la marca de vehículo que ocupa electricidad en el estado de Washington.

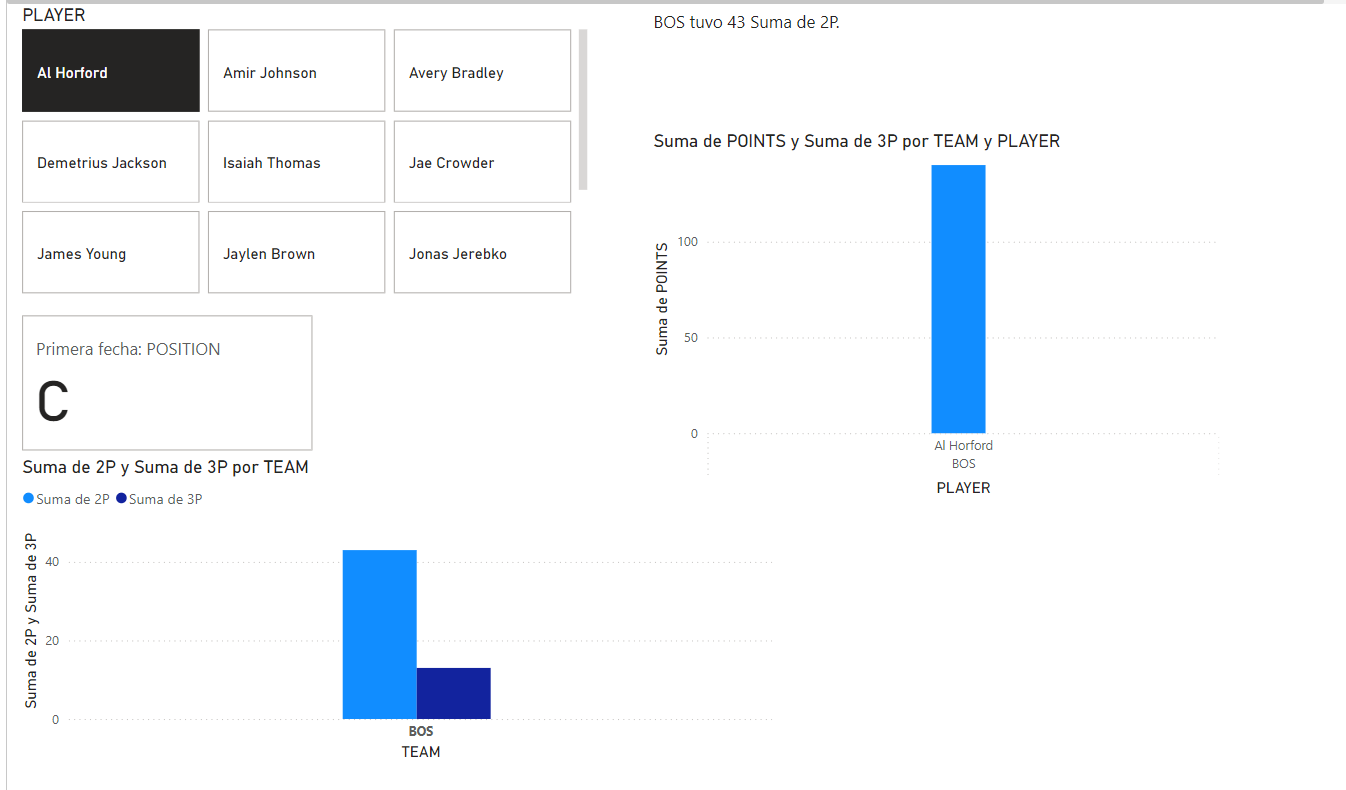


Juego de loteria

Se puede visualizar el conteo de una serie de personas que ganaron



Baloncesto



* Resultados obtenidos.

Covid

<https://app.powerbi.com/view?r=eyJrIjoiMzQzNjZhOTEtM2FjNi00ODcyLWI1MzItODQzNmEwNzEzNzgyIiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9>

Futbuol

<https://app.powerbi.com/view?r=eyJrIjoiYzFlMDg1NDItZjBjMy00NmM5LWE4NjMtN2E4ZTEwYjQ0MzAxIiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9>

Game

<https://app.powerbi.com/view?r=eyJrIjoiZjkxNjFkOTgtZmE5MS00MzI2LTk4ZGEtM2IwMDJmZTE2MGEyIiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9>

Empresa Ventas

<https://app.powerbi.com/view?r=eyJrIjoiYTBhZmZhOTQtMzhjNy00MWJjLTg4MjctYTM3OGRkY2VhNDI3IiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9>

Series de televisión

<https://app.powerbi.com/view?r=eyJrIjoiZjY5ZWI0ZmItZWYyNS00YThjLWFkNWUtMDZhYjNiZjhjZDYwIiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9>

Ventas videojuegos

<https://app.powerbi.com/view?r=eyJrIjoiNGY2NDcwYTUtYmQxNC00Nzg1LTk3MGMtOGQ3NzU3MDM2N2U5IiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9>

Restaurante

<https://app.powerbi.com/view?r=eyJrIjoiZGJiYTE1OWItYmE5NC00YjhkLWIzYjUtMjU2YjZlNzBmZWQxIiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9>

Electricidad acorde a los vehículos en WA-USA

<https://app.powerbi.com/view?r=eyJrIjoiMjFiNmQzZjEtNWYyZi00MjhmLThiODgtMDA2ZTI0ZDc1OGMxIiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9>

Juego de lotería

<https://app.powerbi.com/view?r=eyJrIjoiZGJmMGJkOGYtMWNhNC00NjlhLThkMjMtNWY0OTRkZWVlNzA2IiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9>

Baloncesto

<https://app.powerbi.com/view?r=eyJrIjoiN2ViN2JjMmUtNGUyZC00Yzk4LTk2YzQtYWVjMzgyY2RlMGNiIiwidCI6IjY4MmE0ZTZhLWE3N2YtNDk1OC1hM2FjLTllMjY2ZDE4YWEzNyIsImMiOjR9>

Conclusiones y Recomendaciones.

Integración Multibase de Datos: El proyecto demuestra la capacidad de integrar diferentes tipos de bases de datos, tanto relacionales como NoSQL, en un entorno cohesionado. Esta integración es crucial para manejar datasets diversos provenientes de múltiples fuentes, permitiendo un almacenamiento centralizado en un repositorio único. Esta capacidad es fundamental en entornos empresariales modernos, donde la diversidad de formatos y orígenes de datos es la norma.

Interoperabilidad Eficiente: El uso de MySQL como intermediario para conectar bases de datos SQL y NoSQL muestra una solución efectiva para la interoperabilidad entre diferentes sistemas de gestión de bases de datos. Esto resalta la importancia de contar con herramientas flexibles que puedan servir como puentes entre tecnologías dispares, facilitando el flujo de datos en una infraestructura compleja.

Centralización y Análisis de Datos: La creación de un repositorio central en SQL-Server permite consolidar los datos para un análisis más eficiente. Esto facilita la creación de dashboards y la realización de consultas temáticas mediante Power BI, proporcionando a los usuarios finales una visión unificada y accesible de la información.

Optimización del Flujo de Datos: La optimización del flujo de datos entre las diferentes bases de datos garantiza que la integridad y la eficiencia en la transferencia de información se mantengan, lo cual es esencial para evitar inconsistencias y mejorar el rendimiento general del sistema

### Recomendaciones

Monitoreo Continuo del Rendimiento: Se recomienda implementar herramientas de monitoreo continuo para supervisar el rendimiento de la integración entre las bases de datos. Esto ayudará a identificar y solucionar cuellos de botella o problemas de latencia en tiempo real, asegurando que la infraestructura funcione de manera óptima.

Seguridad y Control de Acceso: Dada la naturaleza distribuida del sistema, es fundamental implementar políticas de seguridad robustas. Esto incluye la encriptación de datos en tránsito y en reposo, así como un control de acceso granular para proteger la información sensible.

Documentación y Mantenimiento: Es crucial mantener una documentación detallada de toda la infraestructura y los procesos de integración. Esto facilitará el mantenimiento del sistema a largo plazo y permitirá a nuevos miembros del equipo comprender rápidamente el entorno.

Pruebas y Validaciones Regulares: Realizar pruebas periódicas de la integridad de los datos y la eficiencia de las transferencias entre bases de datos es esencial para asegurar que el sistema esté funcionando correctamente. Esto puede incluir pruebas de carga y validaciones de consistencia de datos.

Exploración de Tecnologías Emergentes: Se recomienda explorar nuevas tecnologías y herramientas que puedan mejorar la integración y análisis de datos. Esto podría incluir el uso de plataformas de integración de datos en la nube o la implementación de técnicas avanzadas de análisis de datos, como el aprendizaje automático, para extraer insights más profundos de los datos consolidados.

* Desafíos y problemas encontrados.

Al momento de descargar redistudio solo se puede descargar la forma gratuita y es necesario descargar pagar para descargar el formato premiun que no ayuda a importar y exportar archivo con mayor tamaño y sin error al momento de descarga

Otro problema al momento de transferir datos mediante redis a mongodb lo datos como esta convertido en formato de redis al exportar los datos a mongodb solo se convierte en un solo dato

### URL del Repositorio de GitHub (25%)

<https://github.com/Richard-Robalino/Proyectoanalisis.git>