DATE 05/24/2024

PRUEBA TÉCNICA

Presentación Resultados

Candidato: Diana Catalina Lopera

LinkedIn: https://www.linkedin.com/in/dianaclopera/ **PortFolio:** https://catalinaloperab.webflow.io/about **GitHub:** https://github.com/catalinalopera

CONTENIDO



TEST DE CONOCIMIENTOS

Descripción de Piezas calves dentro el ecosistema Azure.



CASO PRÁCTICO MODELADO DE DATOS

Presentación del diagrama modelo Entidad Relación, diagrama RDV y diagrama BDV.



CASO PRÁCTICO ETL CON PYTHON

Explicación de los pasos llevados a cabo para crear el caso prático, además de la presentación del plan de tareas.



DUDAS Y PREGUNTAS

01 TEST DE CONOCIMIENTOS

1.Data Lake

- Organización de los datos: <u>Azure Data Lake Storage (ADLS</u>), porque proporciona un almacenamiento escalable y seguro para grandes volúmenes de datos estructurados y no estructurados en su formato original.
- Gestión de datos sensibles: <u>Azure Data Lake Storage + Azure Key Vault: ADLS</u> porque permite la encriptación de datos en reposo y en tránsito, mientras que Azure Key Vault gestiona de forma segura las claves de encriptación y secretos, asegurando la protección de datos sensibles.
- Movimientos entre capas: <u>Azure Data Factory (ADF)</u> porque facilita la orquestación y automatización de flujos de datos entre diferentes capas del Data Lake (ingestión, almacenamiento crudo, procesamiento, presentación). Permite integrar datos de múltiples fuentes y moverlos de manera eficiente.

2.Data warehouse

- Definición de data warehouse y sus componentes principales: <u>Azure Synapse Analytics</u> ya que s una solución integrada que combina almacenamiento de datos y capacidades analíticas, permitiendo la ingesta, preparación, manejo y servicio de datos para análisis.
- Diferencias entre data warehouse, data mart y data lake: <u>Azure Synapse Analytics + Azure Data Lake Storage</u>, teniendo en cuenta que Azure Synapse se utiliza para data warehousing y data marts, proporcionando análisis rápidos y consultas complejas, mientras que Azure Data Lake Storage es ideal para almacenar grandes volúmenes de datos en bruto y no estructurados.
- Modelos de datos dimensionales y relacionales: <u>Azure SQL Database</u> es adecuado para implementar tanto modelos de datos dimensionales como relacionales, proporcionando una base de datos relacional escalable y de alto rendimiento.
- ETL (Extracción, Transformación y Carga) y sus fases: <u>Azure Data Factory</u> es eficiente para los procesos ETL, permitiendo extraer datos de diversas fuentes, transformarlos mediante flujos de datos y cargarlos en destinos como Azure Synapse Analytics o Azure SQL Database.

3. Modelado de datos

- Entidades, atributos y relaciones en el modelado de datos: <u>Azure SQL Database</u> es ideal para definir y gestionar entidades, atributos y relaciones en el modelado de datos, manteniendo la integridad referencial y optimizando el rendimiento de las consultas.
- Diagramas de entidad-relación (E-R) y notación Crow's Foot: <u>Azure Data Studio</u> es una herramienta adecuada para crear diagramas E-R y utilizar notación Crow's Foot, proporcionando capacidades de diseño y visualización de esquemas de bases de datos, existen varias extensiones que se pueden integrar.
- Normalización de datos y sus diferentes niveles: <u>Azure SQL Database</u> proporciona un entorno robusto para implementar la normalización de datos en diferentes niveles, asegurando la reducción de redundancias y mejora de la integridad de los datos.
- Diseño de dimensiones y hechos en un data warehouse: <u>Azure Synapse Analytics</u> usado para diseñar y almacenar dimensiones y hechos, optimizando el análisis y las consultas de datos.
- Componentes básicos del modelado Data Vault: <u>Azure Synapse Analytics</u> soporta el modelado Data Vault, permitiendo la implementación de hubs, links y satélites, y proporcionando un almacenamiento masivo y escalable para datos empresariales.



4.SQL - Transformaciones (sintaxis)

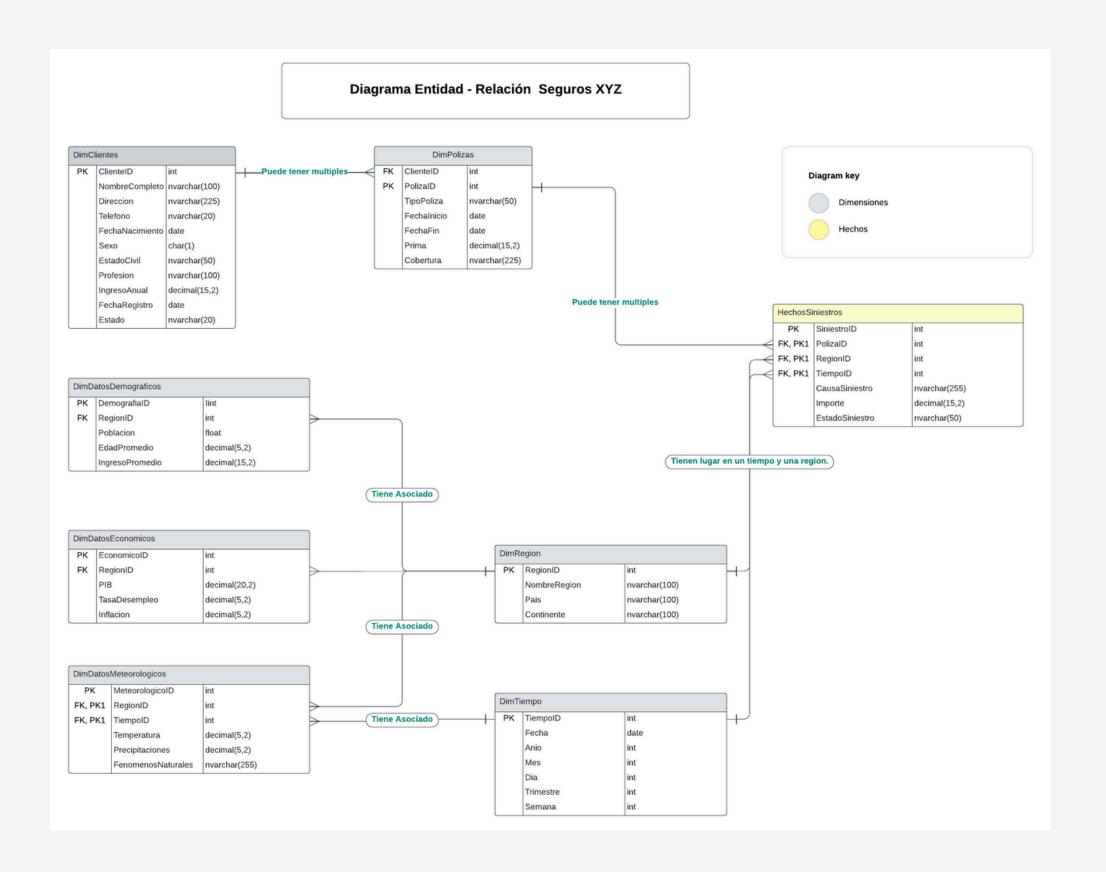
- Consultas básicas de SQL (SELECT, FROM, WHERE, JOIN, ORDER BY, GROUP BY).
- Agregación de datos (SUM, COUNT, AVG, MIN, MAX).
- Subconsultas y correlaciones.
- Funciones de ventana.
- Manipulación de cadenas y fechas en SQL.

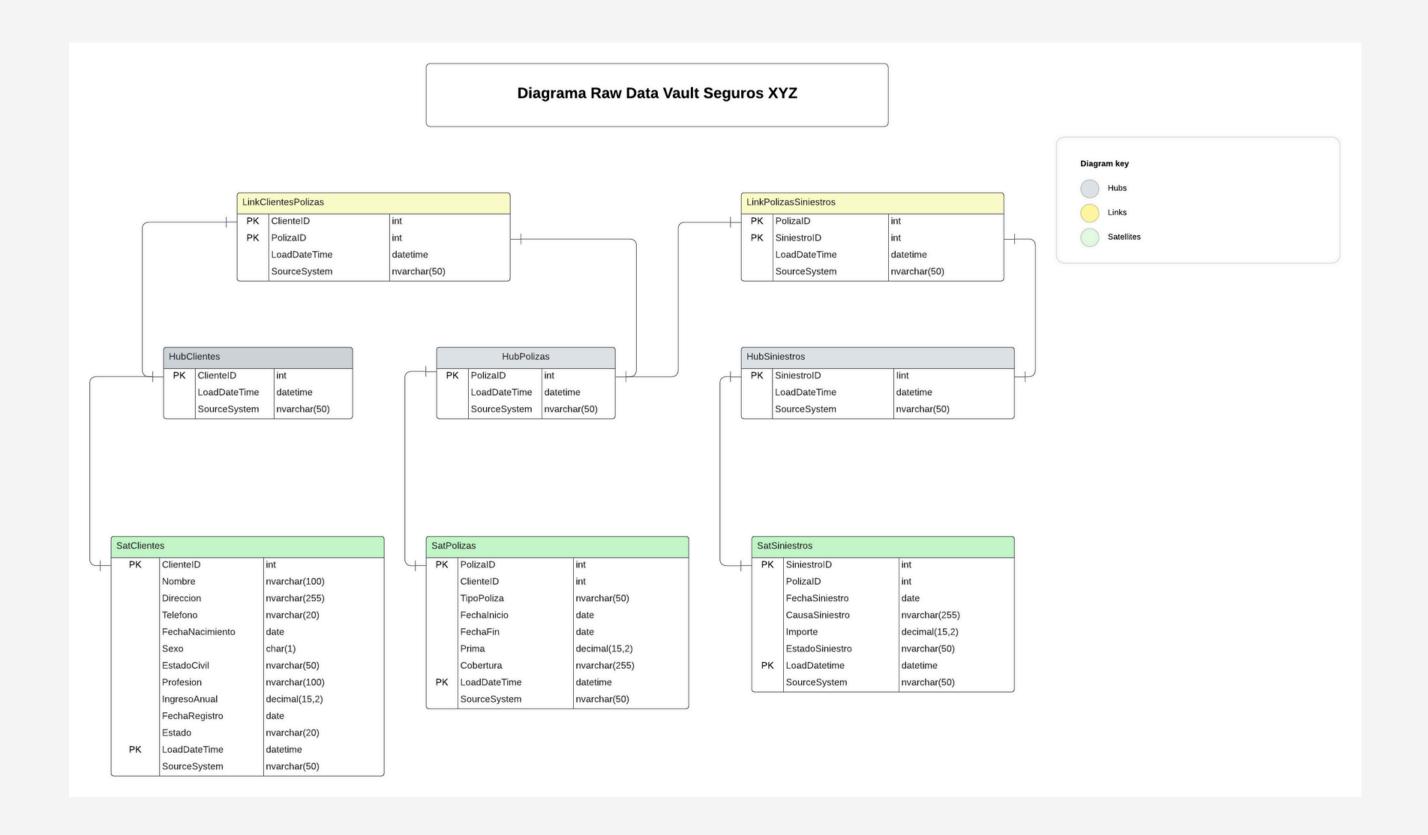
•

<u>Azure SQL Database o Azure Synapse Analytics:</u> Ambas plataformas son adecuadas para ejecutar consultas SQL básicas, ofreciendo un rendimiento robusto y capacidades avanzadas de gestión de datos.



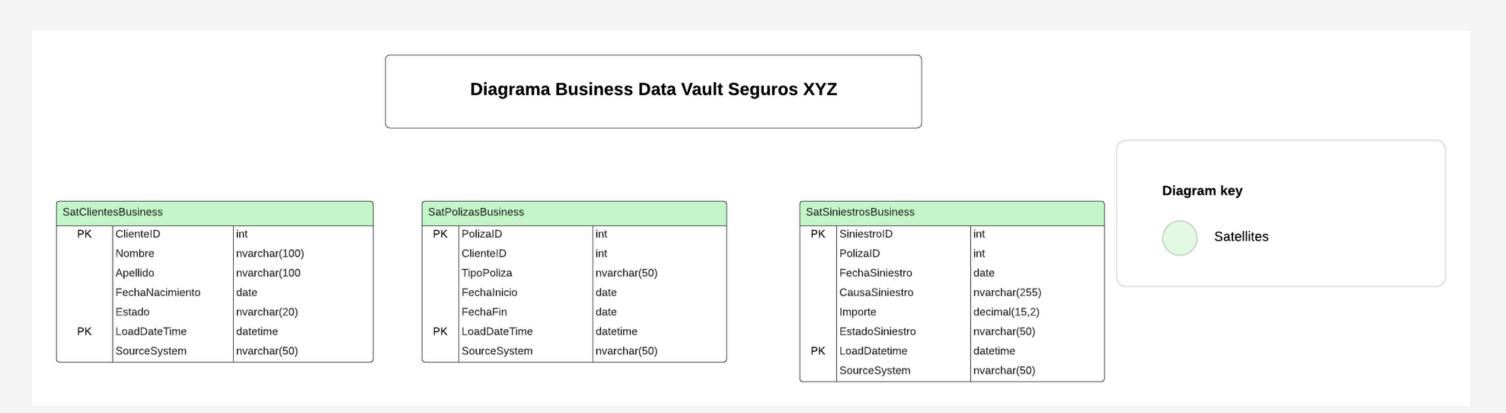
02 CASO PRÁCTICO MODELADO DE DATOS





MODELADO DE DATOS





MODELADO DE DATOS

El Datawarehouse diseñado permite a la empresa de seguros XYZ realizar análisis de datos para:

- Identificar clientes con mayor riesgo de siniestralidad.
- Detectar patrones de fraude.
- Desarrollar nuevos productos y servicios de seguros.
- Mejorar la atención al cliente.

Estadisticas relacionadas con el caso práctico de modelado de datos:

Tiempo Invertido: 2 horas

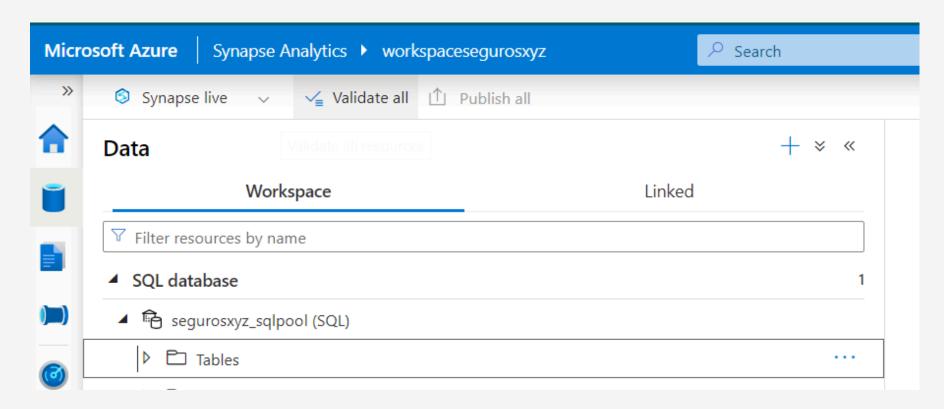
Uso de ChatGPT: Utilicé ChatGPT para aclarar algunas dudas que tenía para modelar el Raw Data

Vault y Business Data Vault.

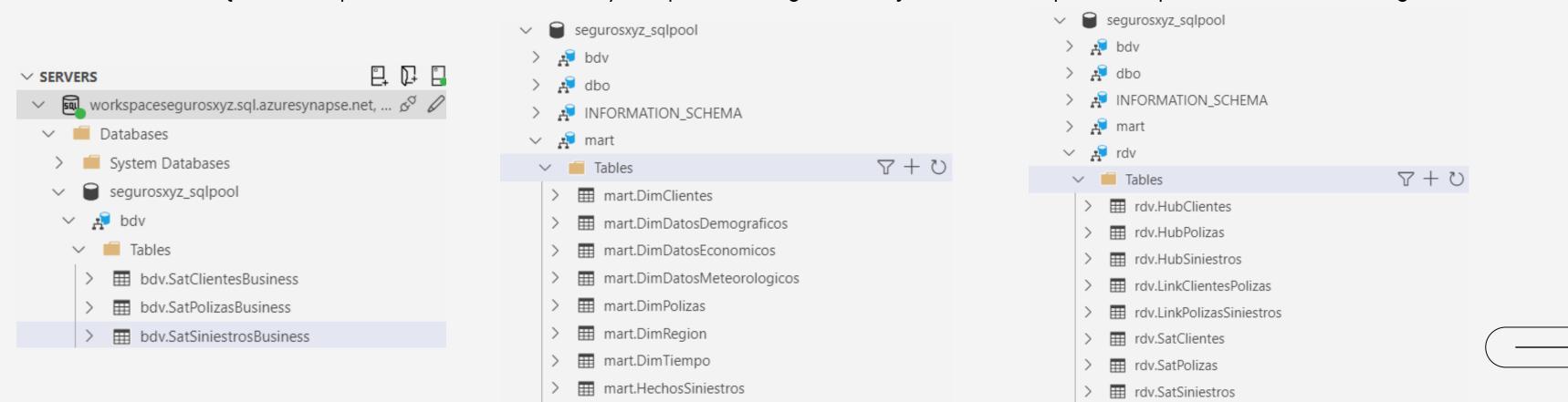
→

03 CASO PRÁCTICO ETL CON PYTHON

1.Se crea un pool de SQL dedicado en Azure Sypnase Analytics, para manipulación de los datos utilicé Azure Data Studio.

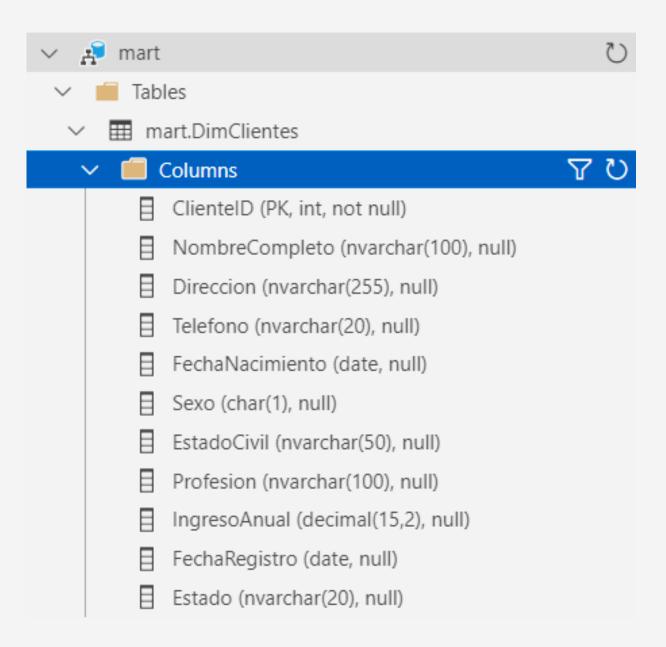


2. Se crean tres esquemas respectivamente mart, bdv y rdv para albergar los objetos creados para cumplir con el caso de negocio.



CASO PRÁCTICO ETL CON PYTHON

1.Se crea tabla DimClientes en el datawarehouse.

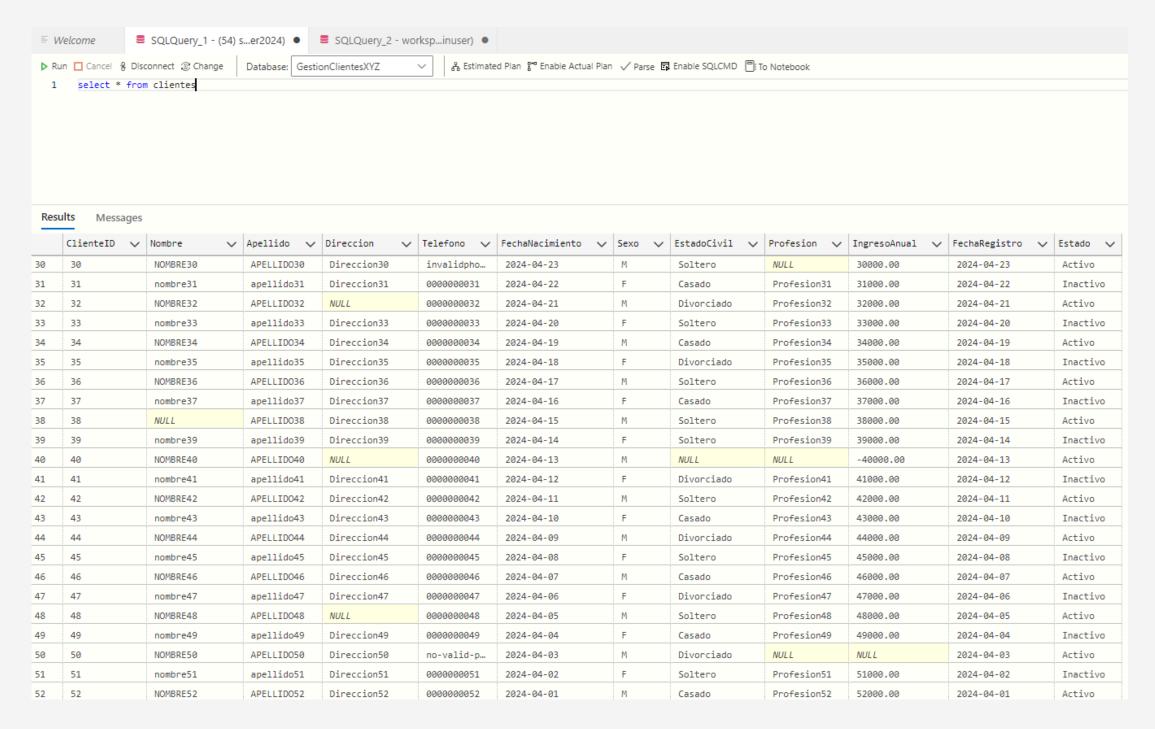


2. Se crea un servidor <u>Azure SQL</u> dentro del mismo resourcegroup de Sypnase Analytics en Azure, alli se crea una base de datos que alberga el sistema de clientes de la compañia XYZ seguros.



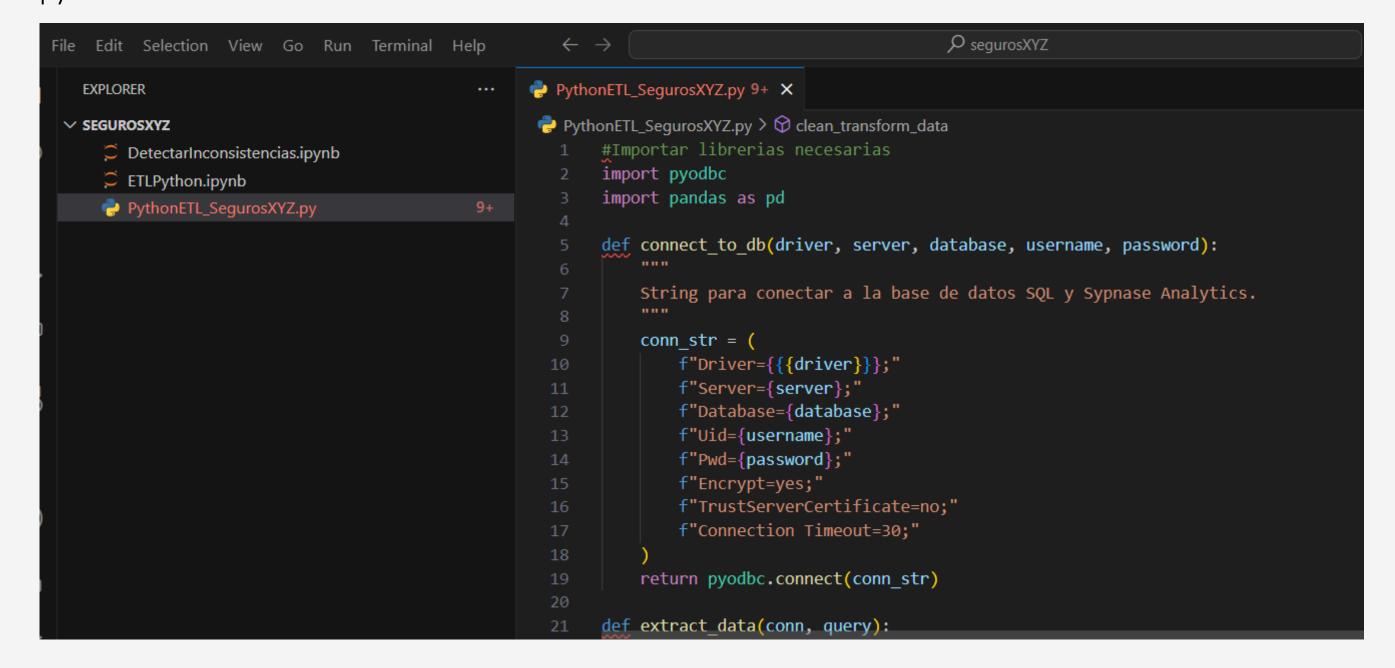


3. Se crea tabla Clientes y se generan 500 registros aleatorios, donde en algunos de ellos se presentan inconsistencias. Por medio de un script de python se resumen las inconsistencias encontradas.



```
Inconsistencias encontradas:
Valores nulos:
ClienteID
                    0
Nombre
                   45
Apellido
                    0
Direccion
                   57
Telefono
                    0
FechaNacimiento
                    0
Sexo
                    0
EstadoCivil
                   23
Profesion
                   43
IngresoAnual
FechaRegistro
                    0
Estado
                    0
dtype: int64
Filas con teléfonos inválidos: 29
Filas con ingresos anuales negativos: 11
Filas duplicadas basadas en ClienteID: 0
```

4. Se crea un ambiente virtual haciendo uso de <u>Anacanda y visual studio code</u>, se instalan las librerias requeridas para la creación del script en python versión 3.8





5. Se crea un script .py con todos los pasos para realizar el proceso ETL solicitado, dentro del archivo se documenta claramente y de manera precisa cada uno de los pasos.

```
PythonETL_SegurosXYZ.py 9+ X
 🥏 PythonETL_SegurosXYZ.py > 🕅 main
      #Importar librerias necesarias
       import pyodbc
       import pandas as pd
       def connect_to_db(driver, server, database, username, password):
           String para conectar a la base de datos SQL y Sypnase Analytics.
           conn str = (
               f"Driver={{{driver}}};"
               f"Server={server};'
               f"Database={database};"
               f"Uid={username};"
               f"Pwd={password};"
               f"Encrypt=yes;"
               f"TrustServerCertificate=no;"
               f"Connection Timeout=30;"
           return pyodbc.connect(conn str)
       def extract_data(conn, query):
           Extraer los datos de la base de datos SQL - Gestion de Clientes Seguros XYZ
           return pd.read sql(query, conn)
       def validate phone(phone):
           Validar que el número teléfonico sea valido
           if phone.isdigit() and len(phone) == 10:
               return phone
           else:
               return "DatoIncorrecto"
```

```
def clean transform data(df):
   Función para Transformar y Limpiar los datos de la fuente de origen para cargarlo al DataWareHouse
   # Borrar Columnas cuyo nombre viene NULL
   df = df[df['Nombre'].notnull()]
   # Verificar los datos que llegan de otras columnas NULL y asignar un nuevo valor como se detalla acontinuación.
   df['Apellido'].fillna('SinDato', inplace=True)
   df['Direccion'].fillna('SinDato', inplace=True)
   df['Telefono'] = df['Telefono'].apply(validate phone)
   df['EstadoCivil'].fillna('SinDato', inplace=True)
   df['Profesion'].fillna('SinDato', inplace=True)
   df['IngresoAnual'].fillna(0, inplace=True)
   df['IngresoAnual'] = pd.to numeric(df['IngresoAnual'], errors='coerce')
   df['IngresoAnual'] = df['IngresoAnual'].apply(lambda x: x if x >= 0 else 0) # Reemplazar los valores Negativos con Cer
   df['Estado'].fillna('SinDato', inplace=True)
   # Normalizar los textos en la base de datos
   df['Nombre'] = df['Nombre'].str.strip().str.title()
   df['Apellido'] = df['Apellido'].str.strip().str.title()
   df['Direccion'] = df['Direccion'].str.strip().str.title()
   df['EstadoCivil'] = df['EstadoCivil'].str.strip().str.capitalize()
   df['Profesion'] = df['Profesion'].str.strip().str.title()
   df['Estado'] = df['Estado'].str.strip().str.capitalize()
   # Crear una nueva columna que contenga la combinación del Nombre + Apellido
   df['NombreCompleto'] = df['Nombre'] + ' ' + df['Apellido']
   df.drop duplicates(subset='ClienteID', keep='last', inplace=True)
   return df
```



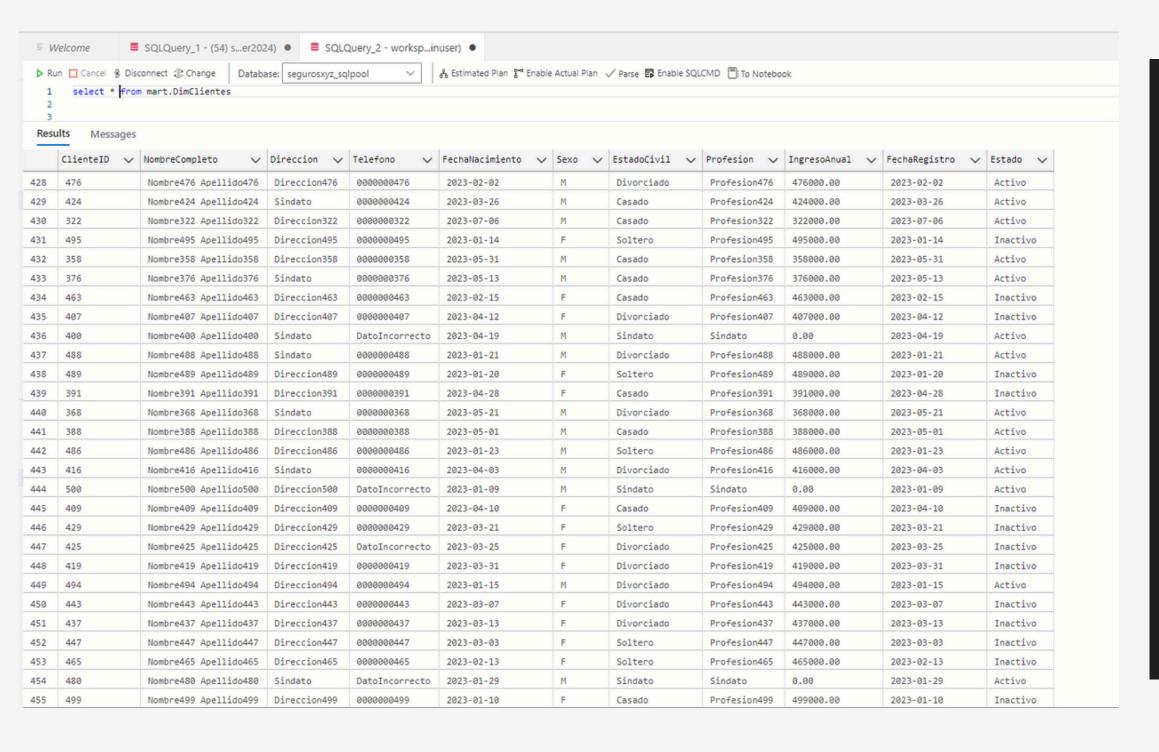
6. Continuación del Script.

```
def load_data(df, conn):
    """
    Función para cargar la data transformada en la tabla del datawarehouse
    """
    cursor = conn.cursor()
    for index, row in df.iterrows():
        cursor.execute("""
        INSERT INTO mart.Dimclientes (ClienteID, NombreCompleto, Direccion, Telefono, FechaNacimiento, Sexo, EstadoCivil, Profesion, IngresoAnual, FechaRegistro, Estado)
        VALUES (?, ?, ?, ?, ?, ?, ?, ?, ?, ?)""",
        row['ClienteID'], row['NombreCompleto'], row['Direccion'], row['Telefono'], row['FechaNacimiento'], row['Sexo'], row['EstadoCivil'], row['Profesion'], row['Ingresconn.commit()
```

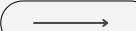
```
def main():
   # Datos de conexión a la base de datos del sistema de gestión de clientes
  client db driver = 'ODBC Driver 18 for SQL Server'
   client_db_server = 'tcp:segurosxyz.database.windows.net,1433'
   client db database = 'GestionClientesXYZ'
   client db username = 'sqladminuser2024'
   client_db_password = 'Segurosxyz.2024'
   # Datos de conexión al data warehouse en Azure Synapse Analytics
   dw driver = 'ODBC Driver 18 for SQL Server'
   dw server = 'tcp:workspacesegurosxyz.sql.azuresynapse.net,1433'
   dw database = 'segurosxyz sqlpool'
   dw username = 'sqladminuser'
   dw password = 'Segurosxyz.2024'
   # Conectar a la base de datos del sistema de gestión de clientes
   conn clientes = connect to db(client db driver, client db server, client db database, client db username, client db password)
   query clientes = "SELECT * FROM Clientes"
   df_clientes = extract_data(conn_clientes, query_clientes)
   # Limpiar y transformar los datos
   df transformed = clean transform data(df clientes)
   # Conectar al data warehouse en Azure Synapse Analytics
   conn_synapse = connect_to_db(dw_driver, dw_server, dw_database, dw_username, dw_password)
   load data(df transformed, conn synapse)
   # Mensaje de confirmación
   print("Los datos se han copiado exitosamente.")
 name == " main ":
  main()
```

CASO PRÁCTICO ETL CON PYTHON

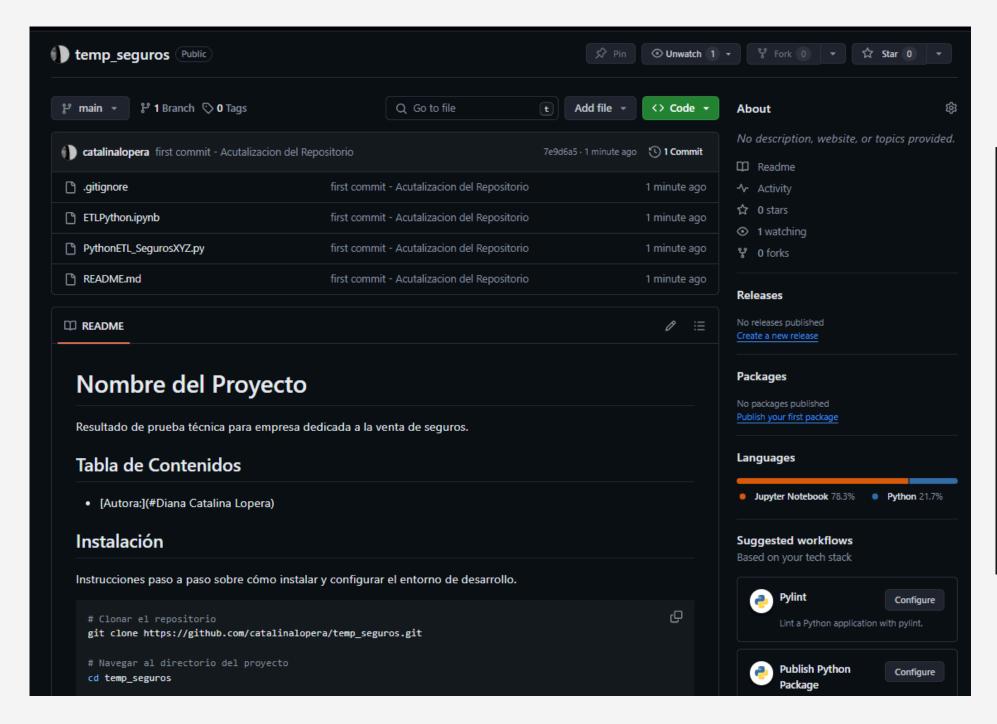
7. Al ejecutarse el Script la tabla DImClientes creada en el datawarehouse en Azure Sypnase Analytics es poblada con un total de 455 registros , y luego de la transaformación de los datos estas son las estadísticas:



```
Inconsistencias encontradas:
Valores nulos:
ClienteID
                   0
NombreCompleto
                   0
Direction
                    0
Telefono
                    0
FechaNacimiento
                   0
Sexo
                    0
EstadoCivil
                   0
Profesion
                   0
IngresoAnual
                    0
FechaRegistro
                    0
Estado
                    0
dtype: int64
Filas con teléfonos marcados como Dato Incorrecto: 29
Filas con ingresos anuales negativos: 0
Filas duplicadas basadas en ClienteID: 0
```



8. Finalmente el script generado es subido a un repositorio de GitHub donde ademas se suben otros archivos utilizados durante la prueba.



PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

PS C:\Users\CATALINA\Desktop\segurosXYZ> git status
On branch main
Your branch is up to date with 'origin/main'.

nothing to commit, working tree clean
PS C:\Users\CATALINA\Desktop\segurosXYZ> []

Repositorio de GitHub: https://github.com/catalinalopera/temp_seguros

9. Construcción del plan de tareas e hitos asociados al desarrollo y evoluciones futuras.

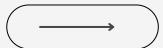
Las tareas pintadas con color naranja hacen parte de la ruta crítica del proyecto de modernización de la infraestructura de datos para mejorar la toma de decisiones y ofrecer un mejor servicio al cliente.

+ Add	Add 📵 Expand all 🗐 Collapse all 🔍 Zoom in 🔍 Zoom out 🐰 Zoom to fit																Search																
IĐ	Name :	Start D	End Date	Duration		May	26, 202	24			Jun 02,	2024			J	Jun 09, 2	2024				Jun 16,	2024				Jun 23,	2024			Ju	n 30, 20	24	4 5
To Halle	Start D	Liid Date	Duration	24	25 26 2	27 28	3 29	30 31	1 1	2 3	4	5 6	7 8	8 9	9 10	11	12 13	14	15 1	6 17	18	19 20	0 21	22 2	3 24	25	26 2	7 28	29 30	1 3	2 3	4	
≣ 1	Entender las fuentes de datos y los requisitos de negocio	May 23, 2024	May 27, 2024	5 days																													
 2	Identificar las tablas relevantes en el sistema de gestión de clientes	May 28, 2024	May 29, 2024	2 days																													
∄ 3	Configurar el entorno de desarrollo en Python	May 30, 2024	May 30, 2024	1 day																													
 4	Instalar las librerías necesarias (pyodbc, pandas, etc.)	May 31, 2024	May 31, 2024	1 day																													
	Extracción: Conectar a la base de datos del sistema de gestión de clientes y extraer los datos	Jun 01, 2024	Jun 03, 2024	3 days																													
 6	Transformación: Limpiar y transformar los datos según las necesidades del data warehouse	Jun 04, 2024	Jun 08, 2024	5 days																													
 	Carga: Conectar al data warehouse y cargar los datos transformados	Jun 09, 2024	Jun 11, 2024	3 days																													
 8	Probar el script ETL con datos reales	Jun 12, 2024	Jun 16, 2024	5 days																													
 9	Validar que los datos se cargan correctamente en el data warehouse	Jun 17, 2024	Jun 19, 2024	3 days																													
∄ 1	Documentar el código y el proceso ETL	Jun 20, 2024	Jun 22, 2024	3 days																													
≣ 1	Crear documentación de usuario y manuales técnicos	Jun 23, 2024	Jun 27, 2024	5 days																													
∄ 1	Planificar la automatización del proceso ETL	Jun 28, 2024	Jun 30, 2024	3 days																													
∄ 1	Identificar mejoras continuas y actualizaciones necesarias	Jul 01, 2024	Jul 05, 2024	5 days																													

Estadisticas relacionadas al caso práctico ETL con Python:

Tiempo Invertido: 3 horas

Uso de ChatGPT: Utilicé ChatGPT para resolver varios inconvenientes que se me presentaron a la hora de crear las conexiones desde Data Studio hacia Azure Sypnase Analytics, asi como tambien para resolver errores presentados durante la transformación de los datos en el script de pyhton.





DUDAS Y PREGUNTAS