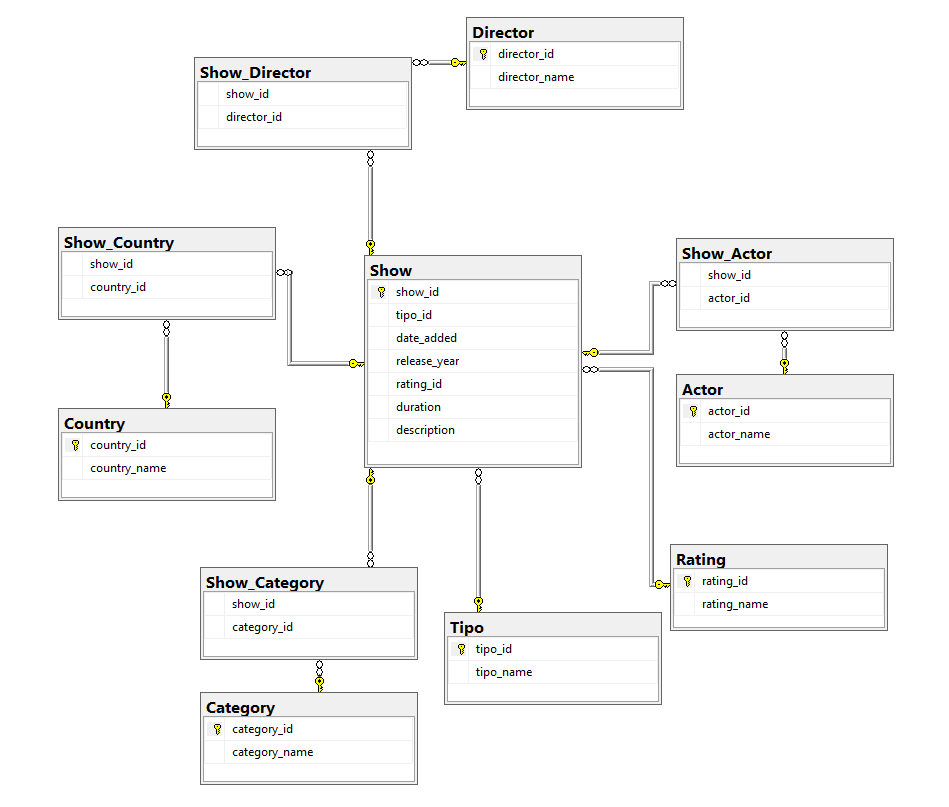
**Desarrollo de la Prueba Técnica**

Postulante: Loaiza Sighuas, Leonardo Dario

1. **Limpieza y Normalización de datos**

Para la realización de esta prueba técnica se empezó descargando la data de Kaggle y normalizando la información a lo cual se llegó al siguiente modelo:



Posteriormente se realizó la limpieza de los datos usando Python. El script de Python se llama “data\_manipulation.py”. Después de esto se realizarían las siguientes acciones:

* Se empezó eliminando los campos nulos usando la biblioteca de pandas. Una vez realizado eso se transformó el campo date\_added a formato datetime para su mejor uso.
* Crear backup de la info original con el campo date\_added modificado y sin campos nulos.
* Separación de columnas cast y creación de nuevo dataframe llamado actores en el que se introducirá solo los datos de la columna cast separados.
* En el nuevo dataframe se eliminan los actores duplicados y se renombran las columnas a actor\_id y actor\_name.
* Lo mismo se hace para las columnas listed\_in, director, y country ya que son columnas donde la relación es de muchos a muchos.
* Posteriormente se obtienen los datos de las columnas rating y type y se les crea su propio dataframe.
* Dividimos la columna duración para quedarnos solo con el dígito.
* Por último, creamos las tablas intermedias entre las hojas y la tabla de hechos (show\_actor, show\_country, show\_category y show\_director), para esto unimos la información de las tablas hojas con la tabla de hechos y dropeamos las columnas que no nos sirven.
* Quitamos las columnas que no nos importan de la tabla de hechos y guardamos los dataframes en archivos csv como backup. Añadimos todos los dataframes a un Excel incluyendo el df original y subimos los datos a la base de datos usando SQLAlchemy (este último paso se hace una vez creada las tablas siguiendo el modelo diseñado de base de datos).

1. **Base de Datos**

Guiándonos del modelo graficado creamos las tablas, la creación de estas se encuentra en el archivo: “create\_query.sql”.

1. **Hoja de Cálculo**

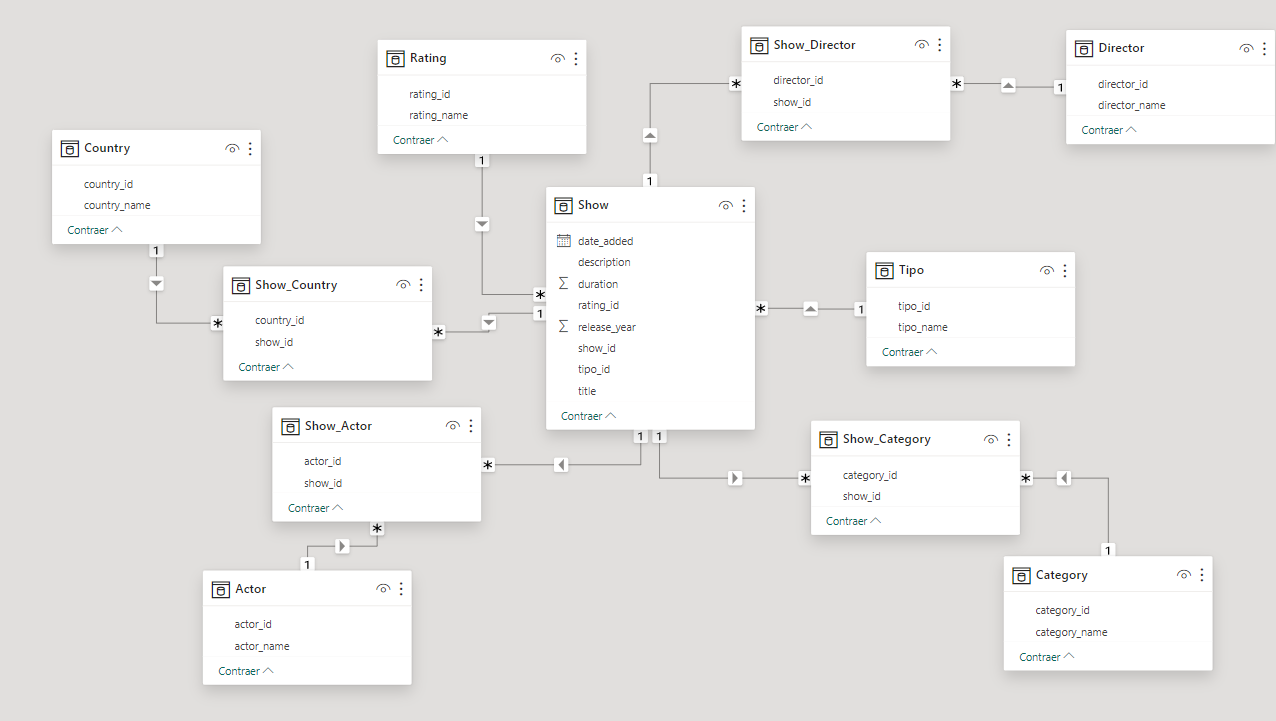
Una vez teniendo los datos cargados podemos crear el reporte. Para esto usé la tabla original procesada y creé 4 gráficos:

* Un gráfico de duración de películas en minutos.
* Un gráfico de barras con la cantidad de shows añadidos por mes.
* Un pie chart de ratings de series de TV.
* Un pie chart de ratings de películas.

Además, también se creó la macro pedida para obtener la cantidad de shows por año teniendo como año la entrada que dé el usuario.

1. **Power BI**

Para crear el dashboard en Power BI primero se importaron los datos de SQL Server.



Una vez importados se eligió el estilo de color de Netflix y se crearon los siguientes gráficos:

* Un mapa que nos permita ver la cantidad de shows por país.
* Un gráfico de anillos que nos permita ver la cantidad de shows por rating.
* Un gráfico de barras verticales de cantidad de shows por director.
* Un gráfico de barras horizontal para ver la cantidad de shows por Actor.
* Un pie chart para ver cantidad de shows por categoría.
* Un gráfico de línea para ver la cantidad de shows por año.
* Un gráfico de barras vertical que nos permita seleccionar o películas o series para ver sus efectos en los demás gráficos.

Una vez terminado el reporte se publicó usando Power BI Service: <https://app.powerbi.com/view?r=eyJrIjoiM2NhY2QxYTctNzE4My00N2UzLWE5NmUtYmZjNjU5Y2I0N2RlIiwidCI6IjM1MWFkNzljLTMyOTQtNGRjNS05Zjg1LTgwZTNjMDI4NzRkYyIsImMiOjR9>