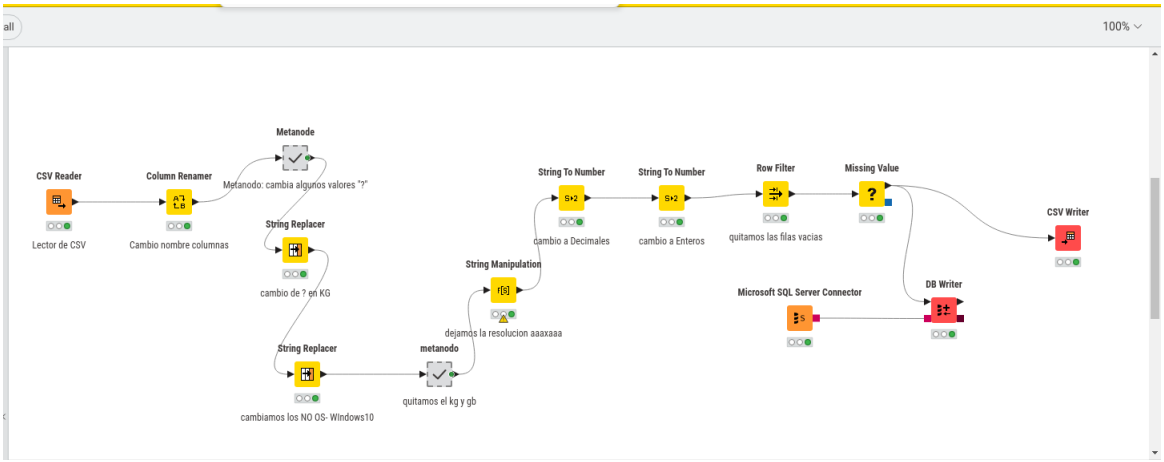
Integrantes:
· Cortés Macias Gretel Penélope 317312184
· Velázquez Barrón Marilú Yatzael 318353492
· Peña Nuñez Axel Yael 318279754
· Escalante Castañeda Lenin Alberto 420003193

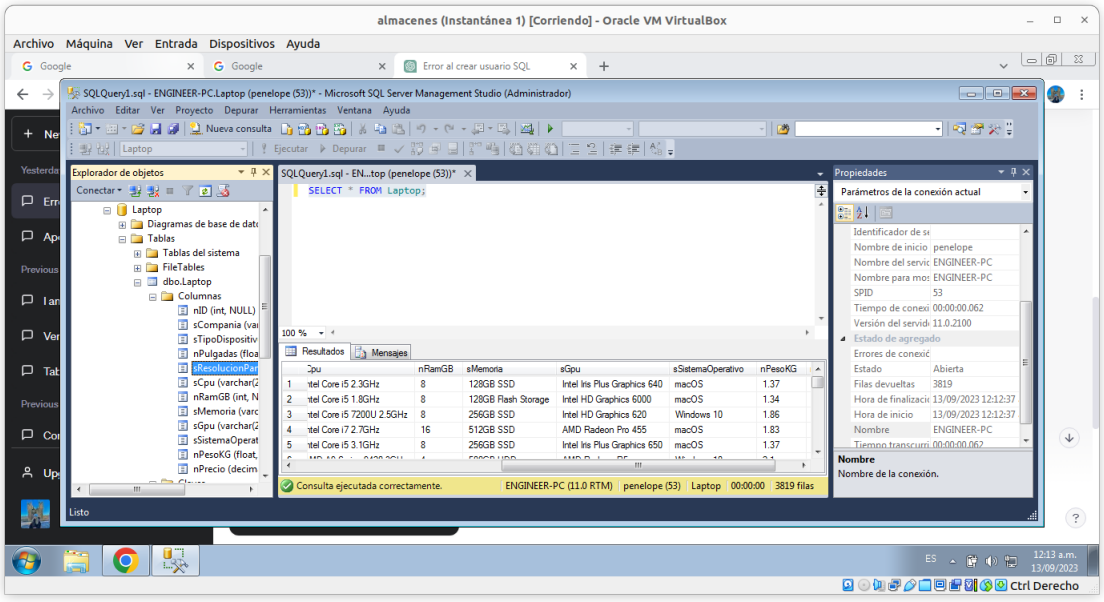
Reporte: KNIME

0.1 KNIME:

En la imagen logramos ver para que nos sirvió cada nodo, incluso hicimos algunos metanodos que nos sirvieron bastante para ver mejor el proceso. No utilizamos diccionarios, lo que creo que nos pudo facilitar algunas acciones, estuvimos usando expresiones regulares que creo que nos complico el trabajo un poco mas. Cambiamos los nombres de las columnas y agregamos un s si es string y una n si es un numero, los string replacer los usamos para 'No OS', o para '?' que aparecia en algunas columnas, el string manipulation nos sirvio para implementar las expresiones regulares donde quitamos KG, GB, y el tipo de la resolucion. Para quitar las filas vacias usamos Row filter y tratamos los missing values con la mediana.



En la parte de conectarnos con la base de datos fue algo confusa porque no lo habiamos hecho, sin embargo fue muy claro el pdf del procedimiento para hacerlo.



0.2 OpenRefine

- **Movies csv**

  Para la columna "Movie", no había mucho que hacer. Se alineó y se quitaron los espacios en blanco del principio y el final. Finalmente, se agrupó usando la opción de "agrupar por huella y colisiones" que el programa proporciona. Además, se notó la repetición de algunas películas. Sin embargo, al no contar con un ID que las identificara como únicas, se llegó a la conclusión de que la base de datos pudo haber sido extraída de una página donde los usuarios dejan comentarios sobre la película e información relacionada.

  La columna "Year" fue la que requirió más esfuerzo. Primero, se alinearon las fechas y se eliminaron los espacios en blanco al inicio y al final. Luego, se limpiaron las fechas del tipo "(fecha -)" para dejarlas como "(fecha)". Lo mismo se hizo con las fechas del tipo "(número)(fecha)" y "(fecha 'lugar de estreno')", utilizando expresiones regulares. Al final, se optó por un formato simple, eliminando los paréntesis "()" y dejando solo la fecha para facilitar su lectura. Además, los registros que no eran fechas y no correspondían al contexto en el que se encontraban, fueron eliminados manualmente. Por ultimo se dicidio separar "Year" en "Año de lanzamiento" y "Año de finalizacion" para las series que estaban resgitradas.

  La columna "Genre" fue similar a "Movie". Se alineó y se quitaron los espacios en blanco al inicio y al final. Luego, se agrupó usando la opción de "agrupar por huella y colisiones" que el programa brinda.

  En "One-Line", simplemente se alinearon los registros, se eliminaron los espacios en blanco, y se revisó si había algún registro para agrupar.

  La siguiente columna que tambien llevo bastante trabajo fue la de "Star", aqui nos dimos cuenta que en los registros no solo se encontraba el nombre de las estrellas si no tambien el/la o l@s directores, es por esto que se decidio dividir esta columna en dos: "Star" y "Director", asi sera mucho mas sencillo buscar la información correspondiente. Esto se logro usando la opción que nos da de dividir para a final en el apartado de editar columna utilizar la opción de remplazar y asi quitar los datos que ya no nos eran necesarios.