

Universidad Nacional Autónoma de México

Facultad de Ciencias, 2024 - 1

Almacenes y Minería de Datos

Tarea 02. Procesos de Extracción, Transformación y Carga

Postgresando eso SQLazos



Integrantes

Adrian Aguilera Moreno	421005200
Marco Antonio Rivera Silva	318183583
Sebastián Alejandro Gutiérrez Medina	318287021
Israel Hernández Dorantes	318206604
Alejandra Ortega García	420002495

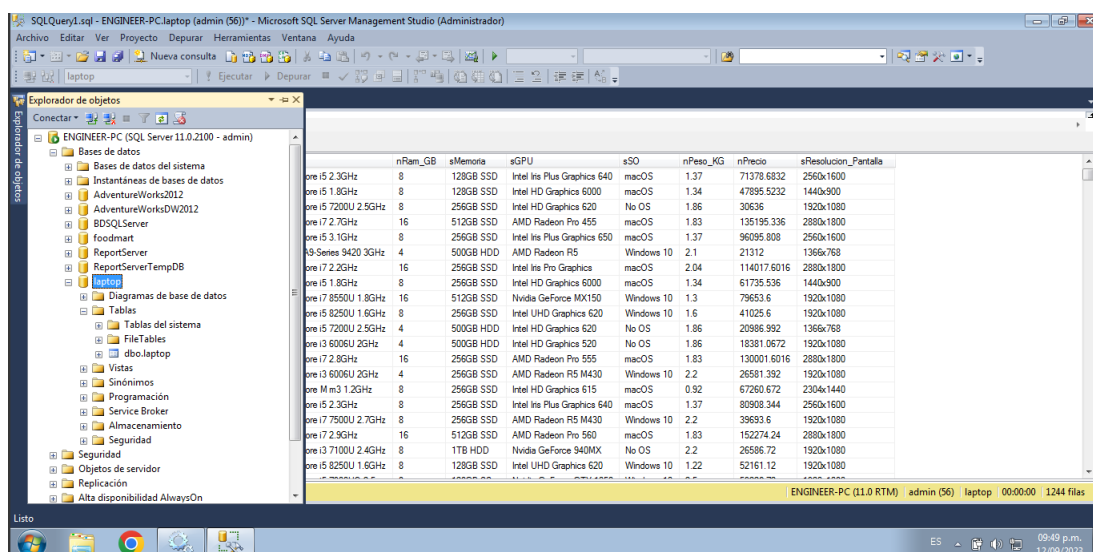
Análisis sobre la limpieza de datos

- **laptopData.csv:**

Para este proceso, seguimos los siguientes pasos:

1. Lo primero que se hizo para la limpieza fue remover las filas vacías, para ello utilizamos un nodo de una librería externa, link del community hub del nodo.
2. Después eliminamos las filas duplicadas sin tomar en cuenta la primera columna con Duplicate Row Filter, pues la primer columna sólo es un enumerador por equipo de laptop.
3. El siguiente paso fue extraer la resolución de la columna ScreenResolution con solo sus números y el caracter "x" que los conecta, para esto se utilizo regex split.
4. Cómo el espacio en RAM puede ser parte de nuestros hechos, entonces decidimos modificar el estilo con el que estaba escrita la columna y eliminamos las terminaciones "GB", esto con la finalidad de volver esta columna a un tipo number.
5. Lo que hicimos después fue estandarizar los formatos de las columnas Memory y OpSys, pues tenían caracteres innecesarios, y agregamos un registro faltante en Weight.
6. Se eliminaron las terminaciones "kg", pues esta información nos resulta valiosa en terminos de number y no tanto cómo un string.
7. Cómo se puede notar, en nuestro renglón 201 en la celda de Weight tenemos un valor "?" que no nos da información acerca del peso del producto, por esto fue sustuido por 2.0 con la ayuda de un nodo Rule Engine.
8. Así, filtramos y renombramos las columnas, pues como resultado de regex split se creo una nueva columna para la resolución limpia, y convertir los tipos las columnas correspondientes a números, a las cuales fueron los datos correspondientes al tamaño de la pantalla, la cantidad de memoria ram y el peso de la laptop.
9. Luego cambiamos los tipos de las columnas; "nPulgadas", "nRam_GB", y "nPeso_KG" a number, pues esta información es más útil en este tipo para generar nuestros hechos.
10. Por último escribimos nuestra información limpia en el archivo.

Obs. Cuando se requiera ejecutar por primera vez el archivo LimpiezaLaptop.knwf, se debe configurar la lectura, en este caso el archivo de dónde lee (laptopData.csv) se encuentra en el directorio Datasets.



	nRam_GB	sMemoria	sGPU	sSO	nPeso_KG	nPrecio	sResolucion_Pantalla
ore i5 2.3GHz	8	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37	71378.6832	2560x1600
ore i5 1.8GHz	8	128GB SSD	Intel HD Graphics 6000	macOS	1.34	47895.5232	1440x900
ore i5 7200U 2.5GHz	8	256GB SSD	Intel HD Graphics 620	No OS	1.86	30636	1920x1080
ore i7 2.7GHz	16	512GB SSD	AMD Radeon Pro 455	macOS	1.83	135195.336	2880x1800
ore i5 3.1GHz	8	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37	96095.808	2560x1600
9-Series 9420 3GHz	4	500GB HDD	AMD Radeon R5	Windows 10	2.1	21312	1366x768
ore i7 2.2GHz	16	256GB SSD	Intel Iris Pro Graphics	macOS	2.04	114017.6016	2880x1800
ore i5 1.8GHz	8	256GB SSD	Intel HD Graphics 6000	macOS	1.34	61735.536	1440x900
ore i7 8550U 1.8GHz	16	512GB SSD	Nvidia GeForce MX150	Windows 10	1.3	79653.6	1920x1080
ore i5 8250U 1.6GHz	8	256GB SSD	Intel UHD Graphics 620	Windows 10	1.6	41025.6	1920x1080
ore i5 7200U 2.5GHz	4	500GB HDD	Intel HD Graphics 620	No OS	1.86	20895.992	1366x768
ore i3 6006U 2GHz	4	500GB HDD	Intel HD Graphics 520	No OS	1.86	18381.0672	1920x1080
ore i7 2.8GHz	16	256GB SSD	AMD Radeon Pro 555	macOS	1.83	130001.6016	2880x1800
ore i3 6006U 2GHz	4	256GB SSD	AMD Radeon R5 M430	Windows 10	2.2	26581.392	1920x1080
ore M m3 1.2GHz	8	256GB SSD	Intel HD Graphics 615	macOS	0.92	67260.672	2304x1440
ore i5 2.3GHz	8	256GB SSD	Intel Iris Plus Graphics 640	macOS	1.37	80908.344	2560x1600
ore i7 7500U 2.7GHz	8	256GB SSD	AMD Radeon R5 M430	Windows 10	2.2	39693.6	1920x1080
ore i7 2.9GHz	16	512GB SSD	AMD Radeon Pro 560	macOS	1.83	152274.24	2880x1800
ore i3 7100U 2.4GHz	8	1TB HDD	Nvidia GeForce 940MX	No OS	2.2	26586.72	1920x1080
ore i5 8250U 1.6GHz	8	128GB SSD	Intel UHD Graphics 620	Windows 10	1.22	52161.12	1920x1080

nId_laptop	sCompania	sTipo	nPulgadas	sCPU	nRam_GB	sMemoria	sGPU	sSO	nPeso_KG	nPrecio	sResolucion_Pantalla
0	Apple	Ultrabook	13.3	Intel Core i5 2.3GHz	8	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37	71378.6832	2560x1600
1	Apple	Ultrabook	13.3	Intel Core i5 1.8GHz	8	128GB SSD	Intel HD Graphics 6000	macOS	1.34	47895.5232	1440x900
2	HP	Notebook	15.6	Intel Core i5 7200U 2.5GHz	8	256GB SSD	Intel HD Graphics 620	No OS	1.86	30636	1920x1080
3	Apple	Ultrabook	15.4	Intel Core i7 2.7GHz	16	512GB SSD	AMD Radeon Pro 455	macOS	1.83	135195.336	2880x1800
4	Apple	Ultrabook	13.3	Intel Core i5 3.1GHz	8	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37	96095.808	2560x1600
5	Acer	Notebook	15.6	AMD A9-Series 9420 3GHz	4	500GB HDD	AMD Radeon R5	Windows 10	2.1	21312	1366x768
6	Apple	Ultrabook	15.4	Intel Core i7 2.2GHz	16	256GB SSD	Intel Iris Pro Graphics	macOS	2.04	114017.6016	2880x1800
7	Apple	Ultrabook	13.3	Intel Core i5 1.8GHz	8	256GB SSD	Intel HD Graphics 6000	macOS	1.34	61735.536	1440x900
8	Aus	Ultrabook	14	Intel Core i7 8550U 1.8GHz	16	512GB SSD	Nvidia GeForce MX150	Windows 10	1.3	79653.6	1920x1080
9	Acer	Ultrabook	14	Intel Core i5 8250U 1.6GHz	8	256GB SSD	Intel UHD Graphics 620	Windows 10	1.6	41025.6	1920x1080
10	HP	Notebook	15.6	Intel Core i5 7200U 2.5GHz	4	500GB HDD	Intel HD Graphics 620	No OS	1.86	20986.992	1366x768
11	HP	Notebook	15.6	Intel Core i3 6005U 2GHz	4	500GB HDD	Intel HD Graphics 520	No OS	1.86	18381.0672	1920x1080
12	Apple	Ultrabook	15.4	Intel Core i7 2.8GHz	16	256GB SSD	AMD Radeon Pro 555	macOS	1.83	130001.6016	2880x1800
13	Dell	Notebook	15.6	Intel Core i3 6005U 2GHz	4	256GB SSD	AMD Radeon R5 M430	Windows 10	2.2	26581.392	1920x1080
14	Apple	Ultrabook	12	Intel Core M m3 1.2GHz	8	256GB SSD	Intel HD Graphics 615	macOS	0.92	67260.672	2304x1440
15	Apple	Ultrabook	13.3	Intel Core i5 2.3GHz	8	256GB SSD	Intel Iris Plus Graphics 640	macOS	1.37	80908.344	2560x1600
16	Dell	Notebook	15.6	Intel Core i7 7500U 2.7GHz	8	256GB SSD	AMD Radeon R5 M430	Windows 10	2.2	39693.6	1920x1080
17	Apple	Ultrabook	15.4	Intel Core i7 2.9GHz	16	512GB SSD	AMD Radeon Pro 560	macOS	1.83	152274.24	2880x1800
18	Lenovo	Notebook	15.6	Intel Core i3 7100U 2.4GHz	8	1TB HDD	Nvidia GeForce 940MX	No OS	2.2	26585.72	1920x1080
19	Dell	Ultrabook	13.3	Intel Core i5 8250U 1.6GHz	8	128GB SSD	Intel UHD Graphics 620	Windows 10	1.22	52161.12	1920x1080
20											

- **movies.csv:**

Para realizar el limpiado de datos con **movies.csv** lo primero que notamos fue que existían muchos espacios en blanco entre los datos, los nombres de las columnas no eran los mejores descriptivos, existían valores "vacíos" en donde debería haber un valor que indicará que no se conocía el dato, en la columna "YEAR" había datos como "(1999 TV Movie)"; y existían celdas donde los años venían de la forma "(2021-)" y otras de la forma "(2021)":

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime	Gross
1	Blood Red Sky	(2021)	Action, Horror, Thriller	6.1	A woman with a mysterious illness is forced into action when a group of terrorists attempt to hijack a transatlantic overnight flight.	Director: Peter Thorwarth Stars: Frit Baumann, Carl Anton Koch, Alexander Scherer, Ralf Selt	21,062	121	
2	Masters of the Universe: Revelation	(2021-)	Animation, Action, Adventure	5.0	The war for Eternia begins again in what may be the final battle between He-Man and Skeletor. A new animated series from writer-director Kevin Smith.	Stars: Chris Wood, Sarah Michelle Gellar, Lena Headey, Mark Hamill	17,870	25	
3	The Walking Dead	(2010-2022)	Drama, Horror, Thriller	8.2	Sheriff Deputy Rick Grimes wakes up from a coma to learn the world is in ruins and must lead a group of survivors to stay alive.	Stars: Andrew Lincoln, Norman Reedus, Melissa McBride, Lauren Cohan	885,805	44	
4	Rick and Morty	(2013-)	Animation, Adventure, Comedy	9.2	An animated series that follows the exploits of a super scientist and his not-so-bright grandson.	Stars: Justin Roiland, Chris Parnell, Spencer Grammer, Sarah Chalke	414,849	23	
5	Army of Thieves	(2021)	Action, Crime, Horror		A prequel, set before the events of Army of the Dead, which focuses on German safecracker Ludwig Dieter leading a group of aspiring thieves on a top secret heist during the early stages of the zombie apocalypse.	Director: Matthias Schwenghofer Stars: Matthias Schwenghofer, Nathalie Emmanuel, Ruby O. Fien, Stuart Martin			
6	Outer Banks	(2020-)	Action, Crime, Drama	7.6	A group of teenagers from the wrong side of the tracks stumble upon a treasure map that unveils a long buried secret.	Stars: Chase Stokes, Madison Cline, Madison Bailey, Jonathan Davis	25,858	50	
7	The Last Letter from Your Lover	(2021)	Drama, Romance	6.8	A pair of interwoven stories set in the past and present follow an ambitious journalist determined to solve the mystery of a forbidden love affair at the center of a love of secret love letters from 1965.	Director: Augustine Frizzell Stars: Shailene Woodley, Joe Alyn, Wendy Nottingham, Felicity Jones	5,283	110	
8	Dexter	(2006-2013)		8.6			665,367	53	

Por lo que cambiamos las celdas que estuvieran vacías por valores como "Unknown" o \$0 según los valores que se manejaban en la columna; también cambiamos los nombres de las columnas para describir el tipo de dato que se usa en cada columna, por ejemplo en "sMovie" en la que la 's' significa que se manejan cadenas de texto (string); modificamos el formato de los años por las inconsistencias que

mencionamos que existían; y a través de las funcionalidades que ofrece **OpenRefine** eliminamos caracteres especiales que tenían algunos datos como espacios en blanco y agregamos una nueva columna, "sType", para poder especificar el tipo de película, como TV Movie, Video, Standard Movie, TV Special, etc., pues en el apartado de "YEAR" se encontraban algunas celdas de la forma "(2021 – TV Movie)":

OpenRefine moviesClean.csv

Enlace permanente

Abrir...ExportarAyuda

Facetas / Filtros

Desahcer / Rehacer 200 / 200

Usar facetas y filtros

Use las facetas y los filtros para seleccionar subconjuntos de sus datos y trabaje en ellos. Puede encontrar estas opciones en los menús de cada columna.

¿Problemas para comenzar? [Vea los videos de ayuda](#)

9999 filas

Mostrar como: filas registros

Mostrar: 51025501005001000filas

< primera< anterior1< siguiente> última>

	Todo	sMovie	sYear	sType	sGenre	nRating	sDescription	sCrew	sVotes	nRuntime	sGross
1.	Blood Red Sky	(2021)	Standard Movie		Action, Horror, Thriller	6.1	A woman with a mysterious illness is forced into action when a group of terrorists attempt to hijack a transatlantic overnight flight.	Director: Peter Thorwarth Stars: Peri Baumerster, Carl Anton Koch, Alexander Schrey, Kati Seltl	21,062	121	\$0
2.	Masters of the Universe: Revelation	(2021)	Standard Movie		Animation, Action, Adventure	5.0	The war for Eternia begins again in what may be the final battle between He-Man and Skeletor. A new animated series from writer-director Kevin Smith.	Stars: Chris Wood, Sarah Michelle Gellar, Lena Headey, Mark Hamill	17,070	25	\$0
3.	The Walking Dead	(2010–2022)	Standard Movie		Drama, Horror, Thriller	8.2	Sheriff Deputy Rick Grimes wakes up from a coma to learn the world is in ruins and must lead a group of survivors to stay alive.	Stars: Andrew Lincoln, Norman Reedus, Melissa McBride, Lauren Cohan	885,805	44	\$0
4.	Rick and Morty	(2013)	Standard Movie		Animation, Adventure, Comedy	9.2	An animated series that follows the exploits of a super scientist and his not-so-bright grandson.	Stars: Justin Roiland, Chris Parnell, Spencer Grammer, Sarah Chalke	414,849	23	\$0
5.	Army of Thieves	(2021)	Standard Movie		Action, Crime, Horror	Unknown	A prequel, set before the events of Army of the Dead, which focuses on German safecracker Ludwig Dieter leading a group of aspiring thieves on a top secret heist during the early stages of the zombie apocalypse.	Director: Matthias Schweighöfer Stars: Matthias Schweighöfer, Nathalie Emmanuel, Ruby O. Fee, Stuart Martin	Unknown	0	\$0
6.	Outer Banks	(2020)	Standard Movie		Action, Crime, Drama	7.6	A group of teenagers from the wrong side of the tracks stumble upon a treasure map that unearths a long buried secret.	Stars: Chase Stokes, Madelyn Cline, Madison Bailey, Jonathan Davis	25,858	50	\$0
7.	The Last Letter from Your Lover	(2021)	Standard Movie		Drama, Romance	6.8	A pair of interwoven stories set in the past and present follow an ambitious journalist determined to solve the mystery of a forbidden love affair at the center of a trove of secret love letters from 1965.	Director: Augustine Frizzell Stars: Shailene Woodley, Joe Alwyn, Wendy Nottingham, Felicity Jones	5,283	110	\$0
8.	Dexter	(2006–2013)	Standard Movie		Crime, Drama, Mystery	8.6	By day, mild-mannered Dexter is a blood-spatter analyst for the Miami police. But at night, he is a serial killer who only targets other murderers.	Stars: Michael C. Hall, Jennifer Carpenter, David Zayas, James Remar	665,387	53	\$0
9.	Never Have I Ever	(2020)	Standard Movie		Comedy	7.9	The complicated life of a modern-day first generation Indian American teenage girl, inspired by Mindy Kaling's own childhood.	Stars: Mahirvi Ramakrishnan, Poorna Jagannathan, Darren Barnet, John McElroy	34,530	30	\$0
10.	Virgin River	(2019)	Standard Movie		Drama, Romance	7.4	Seeking a fresh start, nurse practitioner Melinda Monroe moves from Los Angeles to a remote Northern California town and is surprised by what and who she finds.	Stars: Alexandra Breckenridge, Martin Henderson, Colin Lawrence, Tim Matheson	27,279	44	\$0