Informe SIES
Proyecto Analítico de Datos

Profesor
Mario Ortiz

Alumno
Johan Órdenes
Leonel Villagra

Fecha
17/11/2020

Índice

En esta investigación se nos pide hacer el proceso de ETL, es decir exportar, transformar y cargar los datos del archivo sies.xlsx. Esto lo hicimos realizando la lectura de los datos entregados, limpieza y eliminación de los datos inservibles, y por último el almacenamiento de los datos en una base de datos en Postgres. Para esto separamos nuestro código en dos fases. La primera para la lectura y guardado de archivos que veremos más adelante, y la segunda para subir netamente dichos archivos a la base de datos.

1. Análisis de los datos

ARANCEL ANUAL	COSTO TITULACION	AÑO_DURACION	DURACION CARRERA FORMAL	NIVEL CARRERA O TIPO DE CARRERA	AÑO MATRÍCULA	TOTAL MATRICULA FEMENINO	TOTAL MATRICULA MASCULINO	TOTAL MATRICULA	MATRICULA DE 1ER AÑO FEMENINO	MATRICULA 1ER AÑO MASCULINO	TOTAL MATRICULA 1ER AÑO	MATRÍCULA - % DE MUNICIPAL	MATRÍCULA - % DE PARTICULAR SUBVENCIONADO	MATRÍCULA - % DE PARTICULAR PAGADO	C. Administración Delegada	AÑO TITULADOS	TITULADOS FEMENINO	TITULADOS MASCULINO	TOTAL TITULADOS	% DE COBERTURA PSU EN MATRICULA 1ER AÑO
1.550.000	161.000	2017	5	Técnica de nivel s	2017	29	79	108	15	35	50	51,4%	47,6%	0,0%	1,0%	2016	0	0	0	-
1.550.000	161.000	2017	5	Técnica de nivel s	2017	35	86	121	16	33	49	51,7%	48,3%	0,0%	0,0%	2016	0	0	0	-
1.550.000	161.000	2017	5	Técnica de nivel s	2017	19	48	67	7	24	31	62,1%	37,9%	0,0%	0,0%	2016	0	0	0	-
1.390.000	36.000	2017	4	Técnica de nivel s	2017	0	0	0	0	0	0	NULL	NULL	NULL	NULL	2016	0	0	0	-
3.434.000	214.200	2017	8	Profesional	2017	42	12	54	35	11	46	15,4%	46,2%	38,5%	0,0%	2016	0	0	0	60% <= X <80%
2.214.000	0	2017	10	Profesional	2017	301	32	333	50	5	55	28,3%	71,1%	0,6%	0,0%	2016	29	1	30	40% <= X <60%
2.162.000	0	2017	10	Profesional	2017	292	50	342	70	17	87	25,4%	68,3%	3,0%	3,3%	2016	27	1	28	80% <= X <100%
2.214.000	0	2017	10	Profesional	2017	50	13	63	44	13	57	16,9%	79,7%	0,0%	3,4%	2016	0	0	0	60% <= X <80%
2.214.000	0	2017	10	Profesional	2017	244	33	277	0	0	0	56,6%	34,5%	4,1%	4,9%	2016	58	6	64	-
2.382.000	0	2017	10	Profesional	2017	235	32	267	76	15	91	18,4%	63,9%	11,4%	6,3%	2016	16	2	18	40% <= X <60%
3.382.924	60.000	2017	10	Profesional	2017	238	33	271	49	5	54	13,9%	78,2%	6,0%	1,9%	2016	38	6	44	0% <= X <=100%
3.160.000	0	2017	10	Profesional	2017	120	71	191	48	32	80	16,5%	59,3%	22,5%	1,6%	2016	21	19	40	60% <= X <80%
2.900.000	0	2017	10	Profesional	2017	0	0	0	0	0	0	NULL	NULL	NULL	NULL	2016	0	0	0	-
1.240.000	50.000	2017	4	Técnica de nivel s	2017	136	0	136	73	0	73	69,4%	25,6%	0,0%	5,0%	2016	18	0	18	-
1.240.000	50.000	2017	4	Técnica de nivel s	2017	93	0	93	39	0	39	68,2%	31,8%	0,0%	0,0%	2016	30	0	30	-
1.240.000	50.000	2017	4	Técnica de nivel s	2017	63	0	63	39	0	39	53,3%	46,7%	0,0%	0,0%	2016	27	0	27	-
1.240.000	50.000	2017	4	Técnica de nivel s	2017	88	0	88	59	0	59	50,0%	44,2%	0,0%	5,8%	2016	25	0	25	-
1.240.000	50.000	2017	4	Técnica de nivel s	2017	66	0	66	31	0	31	72,4%	25,9%	0,0%	1,7%	2016	12	0	12	-
3.883.000	0	2017	8	Profesional	2017	121	80	201	24	21	45	12,5%	30,5%	57,0%	0,0%	2016	8	10	18	60% <= X <80%
1.833.300	26.472	2017	5	Técnica de nivel s	2017	64	28	92	38	15	53	13,4%	59,8%	22,0%	4,9%	2016	8	1	9	-
1.833.300	26.472	2017	5	Técnica de nivel s	2017	71	44	115	40	28	68	22,0%	64,2%	3,7%	10,1%	2016	4	4	8	-
2.496.000	100.000	2017	10	Profesional	2017	299	37	336	65	7	72	49,6%	41,5%	1,8%	7,2%	2016	42	2	44	0% <= X <=100%
2.269.000	325.674	2017	10	Profesional	2017	66	12	78	24	7	31	33,3%	65,3%	1,3%	0,0%	2016	13	0	13	-
2.690.000	325.674	2017	10	Profesional	2017	57	8	65	12	4	16	51,8%	37,5%	10,7%	0,0%	2016	17	0	17	-
2.290.000	325.674	2017	10	Profesional	2017	0	0	0	0	0	0	NULL	NULL	NULL	NULL	2016	0	0	0	-
2.798.000	325.674	2017	10	Profesional	2017	26	32	58	7	5	12	21,8%	67,3%	7,3%	3,6%	2016	2	4	6	-
2.498.000	325.674	2017	10	Profesional	2017	41	62	103	9	21	30	34,0%	60,0%	6,0%	0,0%	2016	0	0	0	-
2.690.000	325.674	2017	10	Profesional	2017	45	44	89	16	21	37	62,5%	27,5%	10,0%	0,0%	2016	2	2	4	-
2.910.000	325.674	2017	10	Profesional	2017	108	80	188	25	26	51	22,5%	67,4%	0,6%	9,6%	2016	17	17	34	60% <= X <80%
4.024.000	464.000	2017	8	Profesional	2017	223	37	260	60	7	67	6,9%	29,4%	63,3%	0,4%	2016	30	6	36	60% <= X <80%
2.993.000	183.000	2017	8	Profesional	2017	0	0	0	0	0	0	NULL	NULL	NULL	NULL	2016	0	0	0	-
3.888.000	214.000	2017	8	Profesional	2017	86	41	127	22	17	39	15,6%	72,1%	11,5%	0,8%	2016	8	1	9	0% <= X <=100%
1.336.000	229.000	2017	5	Técnica de nivel s	2017	176	48	224	57	22	79	41,8%	56,0%	0,0%	2,2%	2016	57	8	65	-
1.320.000	229.000	2017	5	Técnica de nivel s	2017	92	11	103	44	2	46	83,5%	11,0%	5,5%	0,0%	2016	0	0	0	-
1.200.000	229.000	2017	5	Técnica de nivel s	2017	148	26	174	64	15	79	69,1%	30,9%	0,0%	0,0%	2016	0	0	0	-
1.500.000	50.000	2017	5	Técnica de nivel s	2017	173	36	209	62	11	73	58,2%	33,7%	0,0%	8,2%	2016	39	7	46	-
1.500.000	50.000	2017	5	Técnica de nivel s	2017	188	30	218	68	11	79	61,6%	38,4%	0,0%	0,0%	2016	40	3	43	-
1.500.000	50.000	2017	5	Técnica de nivel s	2017	161	19	180	56	5	61	42,2%	47,4%	0,0%	10,4%	2016	78	12	90	-
3.740.033	418.603	2017	10	Profesional	2017	163	160	323	32	37	69	29,2%	68,0%	0,9%	1,9%	2016	41	32	73	0% <= X <=100%
3.238.400	373.456	2017	10	Profesional	2017	256	207	463	60	24	84	37,9%	59,3%	0,7%	2,2%	2016	31	38	69	0% <= X <=100%
3.862.500	429.625	2017	10	Profesional	2017	258	203	461	40	35	75	24,4%	71,0%	1,3%	3,3%	2016	35	34	69	0% <= X <=100%
4.243.660	463.624	2017	10	Profesional	2017	211	179	390	53	21	74	22,7%	70,8%	3,7%	2,9%	2016	32	30	62	0% <= X <=100%
3.246.000	86.000	2017	10	Profesional	2017	103	82	185	21	20	41	36,3%	55,5%	8,2%	0,0%	2016	22	15	37	0% <= X <=100%
3.778.977	60.000	2017	10	Profesional	2017	196	222	418	38	58	96	21,4%	66,5%	9,1%	3,0%	2016	15	10	25	0% <= X <=100%
3.675.197	60.000	2017	10	Profesional	2017	209	73	282	78	24	102	12,2%	69,1%	18,3%	0,4%	2016	0	0	0	60% <= X <80%
1.420.000	107.010	2017	4	Técnica de nivel s	2017	68	24	92	19	7	26	12,8%	67,4%	2,3%	17,4%	2016	15	5	20	-
3.018.555	60.000	2017	8	Profesional	2017	124	368	492	17	47	64	23,4%	66,7%	7,1%	2,7%	2016	28	63	91	0% <= X <=100%
2.252.750	421.000	2017	10	Profesional	2017	114	105	219	27	29	56	22,3%	77,2%	0,5%	0,0%	2016	28	26	54	0% <= X <=100%
2.280.560	421.000	2017	10	Profesional	2017	104	59	163	18	12	30	57,4%	40,7%	0,0%	1,9%	2016	21	5	26	0% <= X <=100%
1.550.000	161.000	2017	5	Técnica de nivel s	2017	229	1	230	110	1	111	44,1%	55,9%	0,0%	0,0%	2016	0	0	0	-

Tabla de datos “sies.xlsx”
Fuente: “Escuela de Ingeniería UCN”

Para la lectura de los datos decidimos marcar de color amarillo las columnas que necesitaban limpieza, marcar de color rojo las columnas que necesitaban eliminarse y las columnas blancas sin marcar, dado que no necesitaban ningún cambio.

Las celdas vacías y erróneas contenían los siguientes N/A, NULL, s/i, -, #N/D, #VALOR!, (vacío). Lo que decidimos hacer en este caso fue:

- Pasar datos tipo texto no conocidos como “nan”
- Pasar datos tipo número no conocidos como “nan”

Para la columna “**% de cobertura psu último año**” decidimos dividirla y crear 2 columnas a partir de ésta, dado que contenía 2 valores diferentes dentro de la misma columna, el porcentaje mínimo y el porcentaje máximo de cobertura del último año.

2. Limpieza de los datos

También añadimos funciones que nos sirven para la limpieza luego de leer los datos de cada columna, definiendo los valores NULL como ‘nan’ que son las siguientes:

- cleanPcobertura(): Limpia los datos de la columna “% de cobertura psu último año”.
  - Transforma los datos ‘-’ a NULL
  - Reemplaza el string ‘% <= X’ < con ‘,’
  - Reemplaza el string ‘=’ con ‘’
  - Corta el último carácter del string ‘%’
  - Divide el string en 2 con el carácter restante ‘,’
  - Retorna 2 valores

- cleanText(): Limpia los datos de las columnas tipo texto.
  - Reemplaza el string ‘s/i’ por NULL
  - Reemplaza el string ‘nan’ por NULL

- cleanPercentages(): Limpia los datos de tipo decimal.
  - Reemplaza el string ‘nan’ por NULL

- cleanDigit(): Limpia los datos de las columnas que contienen dígitos enteros.
  - Reemplaza el string ‘s/i’ por NULL
  - Reemplaza el string ‘nan’ por NULL

- cleanDecimal(): Limpia los datos de las columnas que contienen dígitos flotantes.
  - Reemplaza el string ‘s/i’ por NULL
  - Reemplaza el string ‘nan’ por NULL
  - Reemplaza el string ‘ ’ por NULL
  - Reemplaza el string ‘-’ por NULL

También hubo columnas que no se limpiaron, esto fué porque los datos están 100% funcionales y sin problemas para un posterior análisis y agregación a la base de datos, las columnas son las siguientes:

- CODIGO UNICO DE CARRERA
- AÑO_INFORM
- CODIGO DE INSTITUCIÓN
- AREA DE CONOCIMIENTO
- INSTITUCION
- NOMBRE CARRERA
- REGION
- JORNADA
- SEDE
- DURACION CARRERA FORMAL
- AÑO TITULADOS
- PSU PONDERACION NOTAS EM
- PSU PONDERACION RANKING
- PSU PONDERACION LENGUAJE
- PSU PONDERACION MATEMATICAS
- PSU PONDERACION HISTORIA
- PSU PONDERACION CIENCIAS
- PSU PONDERACION OTROS

Además hubo columnas que eliminamos, dado que no eran representativas y no generaban un cambio al momento de agregarla a la base de datos, esto nos ayuda a que haya un ahorro de cantidad de datos ingresados.

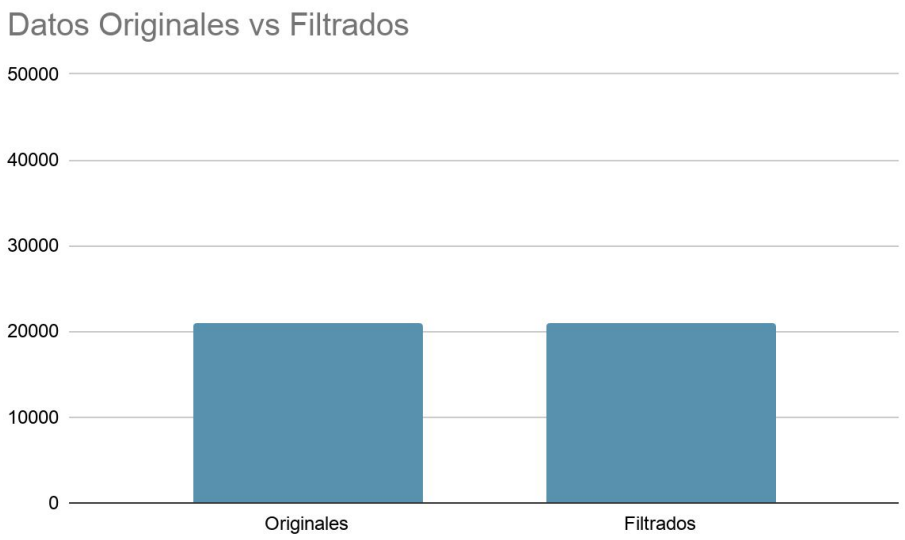
- TIPO DE INSTITUCION: Fué eliminada debido a que todas las filas de esta columna poseían el mismo atributo “universidad”, en este caso estuvimos analizando datos de universidad, entonces no hubo datos de tipo “instututo” o “colegio” por ejemplo.

- AÑO_DURAC: Fué eliminada debido a que ya existía una columna con los mismos datos, solo que tenía otro nombre. En este caso, “Año duracion” contenía los mismos datos de “Año matrícula”.

- TOTAL MATRICULA, TOTAL MATRICULA 1ER AÑO, TOTAL TITULADOS: Estas 3 columnas no eran necesarias ya que por cada una de ellas existen 2, una para la suma de personas femeninas y la otra para la suma de personas masculinas, por ende el total se puede deducir de la suma entre estas 2 columnas.

3. Comparación de datos

Luego de hacer toda la limpieza de datos anterior, decidimos comparar los datos que teníamos al principio, con los datos nuevos dentro del mismo dataframe modificado y llegamos a la siguiente conclusión:



Cantidad de datos Originales vs Cantidad de datos filtrados
Fuente: Elaboración de Equipo

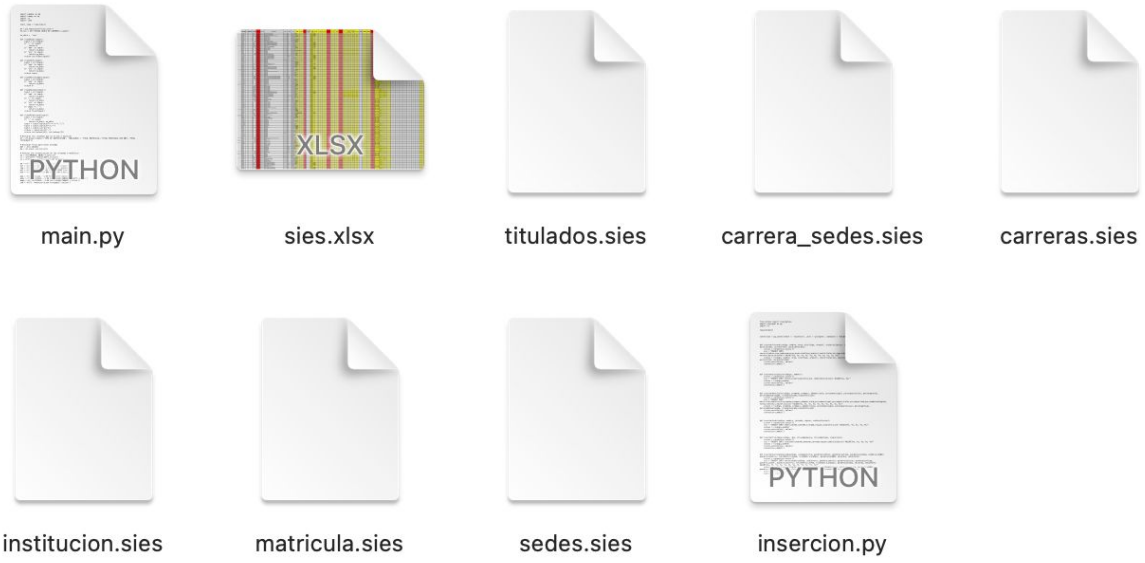
La cantidad de datos originales era de 20.944 columnas, mientras que luego de haber limpiado y filtrado todos los datos quedó la misma cantidad de filas que teníamos al principio 20.944. Esto significa que hubo un 0% de pérdida de información.

4. Tiempo de compilación

El código final se compiló en promedio 11.445152044296265 segundos sin contar la subida a la base de datos, es decir, desde el momento de leer el archivo “sies.xlsx” hasta generar los archivos de texto en formato “.sies” para su posterior subida a la base de datos.

El algoritmo más costoso es de orden O(n), en donde n es la cantidad de ciclos. Estos algoritmos que obtuvieron dicho orden, son los encargados de filtrar los datos y guardar los archivos “.sies” para subir a la base de datos.

5. Exportación de archivos



Archivos del proyecto
Fuente: Elaboración propia

Aquí se tienen todos los archivos que se generan luego de ejecutar el código principal. Se generan los archivos “titulados.sies”, “carrera_sedes.sies”, “carreras.sies”, “institucion.sies”, “matricula.sies”, “sedes.sies” para su posterior carga a la base de datos. Lo hicimos de esta manera (Importación-Exportación) para evitar la espera que se procesen los datos mientras hacemos pruebas al programar. Esto puede mejorar haciendo uso de hilos.

6. Carga a la bases de datos

Para la elección de tablas decidimos crear 6 tablas con las siguientes claves y detalles:

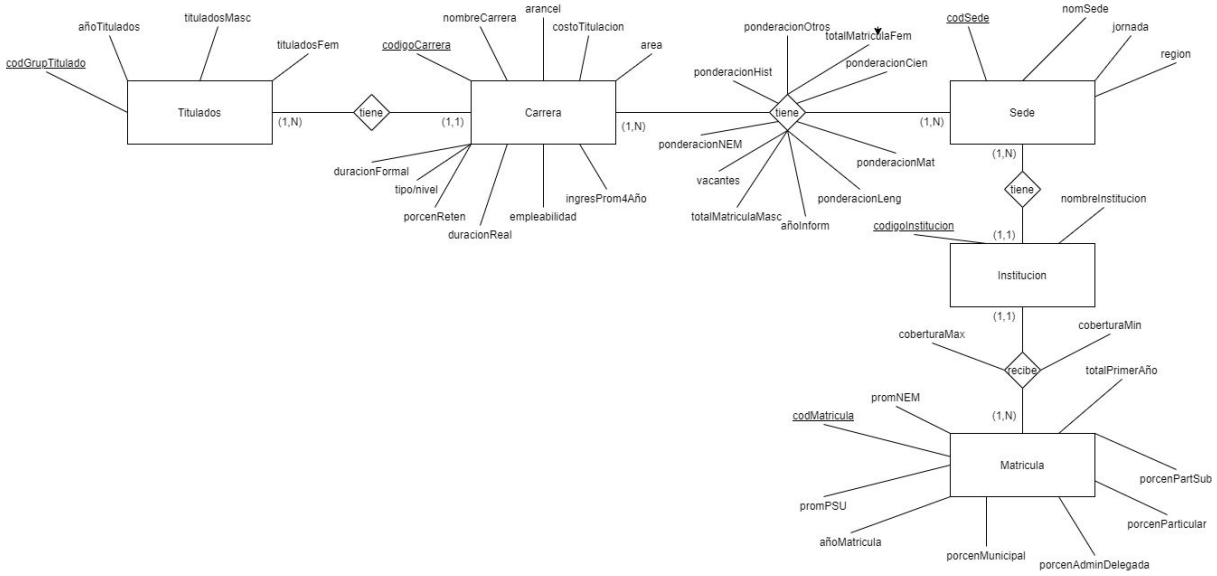


Diagrama Entidad-Relacionamiento
Fuente: Elaboración de Equipo

- Titulado: Llave primaria nueva creada como serializable.

La tabla titulados decidimos dejarla así saber los datos de los titulados por cada carrera, es ilógico dejarlo en matrícula o sede, dado que la cantidad de alumnos incrementa considerablemente.

- Carrera: Código de la carrera como llave primaria.

Una carrera puede estar en muchas sedes y en una sede pueden haber muchos titulados, si fuera por institución o matrícula, no lo hicimos junto a la institución porque cada carrera puede variar por sede, incluso los datos de las ponderaciones PSU.

- Sede: Nombre de la sede como llave primaria.

Una sede está claramente entre una institución y carrera, esto es lo más lógico que se puede sacar del diagrama.

- Carrera-Sede: Carrera y Sede como llave primaria y foránea.

Decidimos agregar una tabla intermedia para las ponderaciones de la PSU entre carrera y sede, no se realizó en matrícula porque las ponderaciones se deben analizar por carrera y la sede a la cual pertenecen.

- Institución: Código de institución como llave primaria.

Esta también es claramente una tabla simple, dado que la institución solo se verifica por el nombre.

- Matrícula: Llave primaria nueva creada como serializable.

Aquí decidimos poner el promedio general de PSU, NEM y los porcentajes, dado que estos están asociados sí o sí a una matrícula y no a una carrera de una sede.