Instituto Politécnico Nacional

Escuela Superior de Cómputo

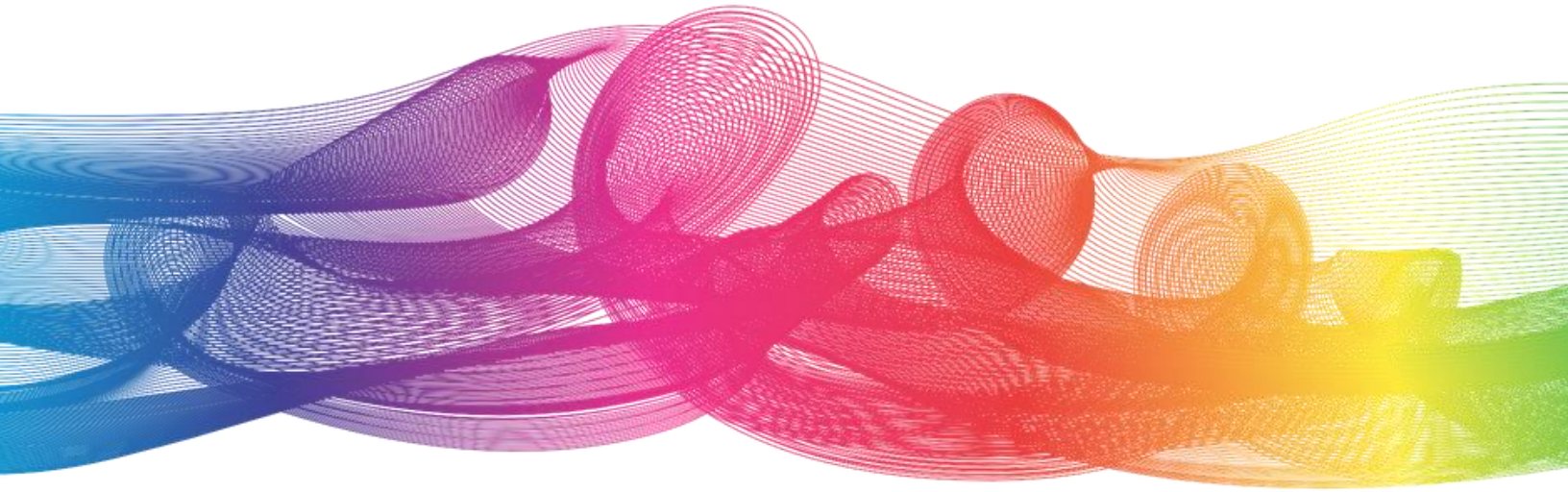
Grupo 3CV19 Data Mining

Profesor Zagal Flores Roberto Eswart

2do Parcial Practica 4: Proceso de ETL, caso "precipitación pluvial en CDMX"

Alumno García González Aarón Antonio

Miercoles 28 de Abril de 2021



Índice

Objetivos

Objetivo: Desarrollar una herramienta ETL para procesar archivos de Excel para el caso “Precipitación pluvial (PP) con la técnica de recolección para depósito húmedo (H)”, durante el periodo “2010 al 2019”.

Introducción

Se propone realizar la herramienta ETL lo mas especifica para esta practica, a su vez realizar cada proceso de la manera mas automatizada posible, por ello después de descargar los archivos zip y descomprimir, se propone buscar los archivos objetivo mediante la terminal de comandos, utilizar Python para la transformación de datos y la librería de pandas para el manejo de los archivos xls y csv, posterior se unen los archivos y se hacen las validaciones y operaciones de limpieza necesarias, se exportan los archivos a analizar, en sql server se importa la tabla de hechos principal y el catalogo de estaciones, ahí se analizaran las tendencias y exploración de datos.

Desarrollo

Procedimiento: Analizar la estructura de los archivos de Excel de origen, y determinar un proceso de integración de datos, que incluya tareas, reglas y/o validaciones.

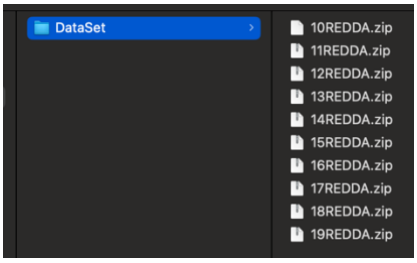
1. Revise los metadatos de la fuente, disponible en: ☑

   http://www.aire.cdmx.gob.mx/descargas/datos/excel/REDDAxls.pdf

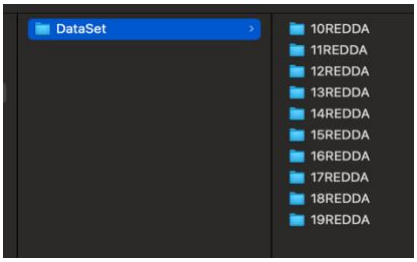
2. Descargue, de la red de depósito atmosférico ☑

   http://www.aire.cdmx.gob.mx/default.php?opc=%27aKBk%27 , los archivos con terminación: "*PPH.XLS", para el periodo del "2010 al 2019".

   Descargar los archivos *REDDA.zip



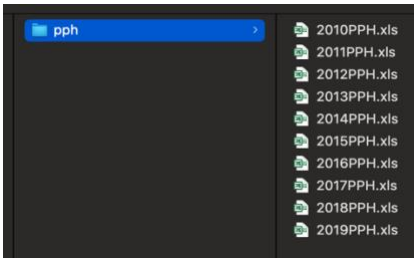
   Descomprimir los archivos *REDDA.zip



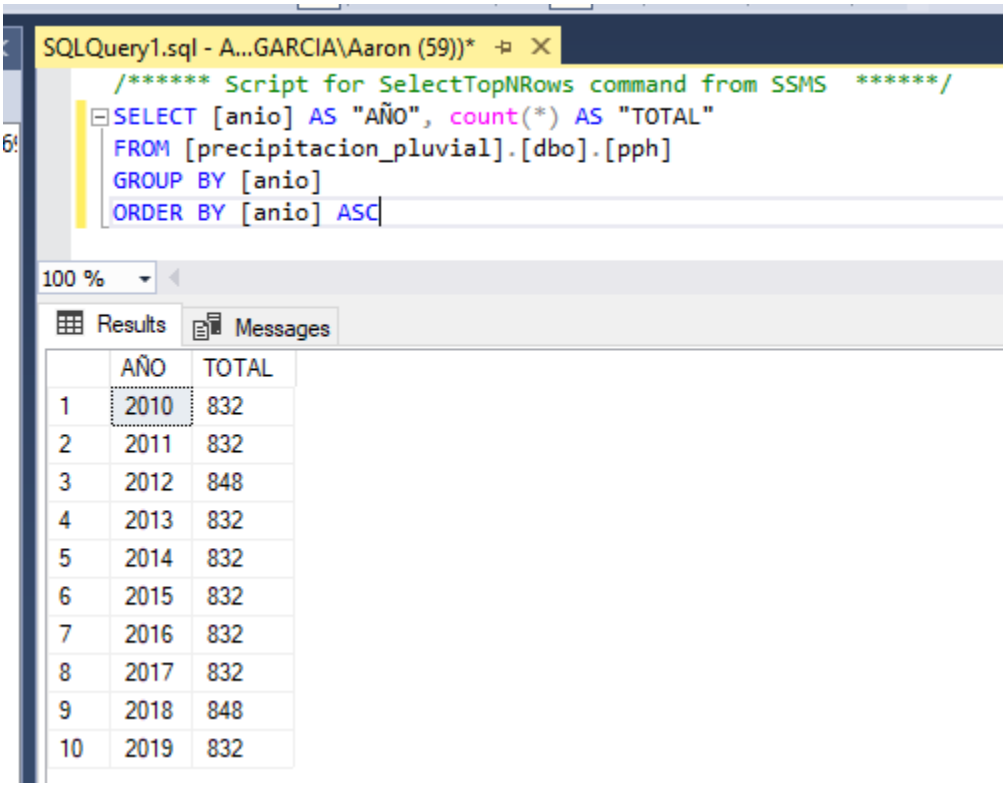
   Creamos una carpeta "pph", buscar en el path indicado los archivos *PPH.xls" y estos los copiamos a la carpeta creada.

```
[aarongarcia@Aarons-MacBook-Pro dataset % mkdir -p /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph & find /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset -name "*PPH.xls*" -exe
c cp "{}" /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph  \;
[1] 50638
[1]  + done       mkdir -p /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph
cp: /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2017PPH.xls and /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2017PPH.xls are identical (not copied).
cp: /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2015PPH.xls and /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2015PPH.xls are identical (not copied).
cp: /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2019PPH.xls and /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2019PPH.xls are identical (not copied).
cp: /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2013PPH.xls and /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2013PPH.xls are identical (not copied).
cp: /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2016PPH.xls and /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2016PPH.xls are identical (not copied).
cp: /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2014PPH.xls and /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2014PPH.xls are identical (not copied).
cp: /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2018PPH.xls and /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2018PPH.xls are identical (not copied).
cp: /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2012PPH.xls and /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph/2012PPH.xls are identical (not copied).
aarongarcia@Aarons-MacBook-Pro dataset %
```

Y listo, tenemos los archivos objetivo: "Precipitación pluvial (PP), recolección para depósito húmedo (H)", en el periodo 2010-2019.



3. Identifique y defina cómo extraer (definición de flujo de integración de datos, reglas y validaciones): ☑
   ⇒ Año,
   ⇒ Elemento de medición, en este caso “Precipitación pluvial (PP)”
   ⇒ La semana de medición, mes y año
   ⇒ La ubicación de la medición
   ⇒ Valor de la medición, en este caso “Precipitación pluvial (PP)”
   ⇒ Defina las reglas necesarias para transformar los datos, por ejemplo, colocar el número de semana del año.
4. Desarrolla la estructura de la tabla de hechos y los respectivos catálogos, todos los que sean necesarios. ☑
   ⇒ Catálogo de estaciones de monitoreo
   ⇒ Catálogo de elementos de medición (en este caso que solo contenga al elemento de medición en estudio junto con toda su información (e.g. unidad de medición)
   ⇒ La tabla de hechos mínimo debe tener la siguiente estructura: {elemento},{localizacion},{noSemana},{anio},{fecha},{medicion}
5. Explore los datos integrados, indique y documente:
   ⇒ Cantidad de registros totales y por año



	AÑO	TOTAL
1	2010	832
2	2011	832
3	2012	848
4	2013	832
5	2014	832
6	2015	832
7	2016	832
8	2017	832
9	2018	848
10	2019	832

⇒ Tendencia de las semanas con mayor y menor cantidad de precipitación pluvial

SQLQuery6.sql - A...GARCIA\Aaron (63))* | SQLQuery5.sql - A...GARCIA\Aaron (56))

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT [week], AVG([medicion]) AS "PICOS"
FROM [precipitacion_pluvial].[dbo].[pph]
WHERE [medicion] != -99
GROUP BY [week]
ORDER BY "PICOS" DESC
```

100 %

Results | Messages

	week	PICOS
1	45	74.9529411764706
2	26	52.9919736842105
3	27	51.9831617647059
4	35	50.7288079470199
5	25	50.5981045751634
6	34	43.536050955414
7	29	42.3451369863014
8	10	38.64375
9	33	37.7242446043165
10	28	36.7173684210526
11	32	35.1289743589744
12	37	35.0183823529412
13	36	34.8466666666667
14	30	32.7812666666667
15	38	31.938940397351
16	14	31.6625
17	2	30.9290322580645
18	31	30.641038961039
19	39	30.4371755725191
20	24	29.3117117117117
21	23	29.0512213740458
22	6	26.275
23	42	25.5206896551724
24	40	23.9484848484848
25	22	23.3437007874016
26	21	22.8254198473282

⇒ Indique los lugares con mayor precipitación durante todo el periodo de estudio

SQLQuery7.sql - A...GARCIA\Aaron (52)) SQLQuery6.sql - A...GARCIA\Aaron (63))* SQLQuery5.sql - A...GARCIA\Aaron (56)) SQLQuery1.sq

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT [Delegación_o_Municipio] AS "Delegación o municipio", AVG([medicion]) AS "Promedio de precipitación pluvial"
FROM [precipitacion_pluvial].[dbo].[pph] INNER JOIN [precipitacion_pluvial].[dbo].[estaciones]
ON [precipitacion_pluvial].[dbo].[pph].[localizacion] = [precipitacion_pluvial].[dbo].[estaciones].[id]
WHERE [medicion] != -99
GROUP BY [Delegación_o_Municipio]
ORDER BY "Promedio de precipitación pluvial" DESC
```

100 %

Results Messages

	Delegación o municipio	Promedio de precipitación pluvial
1	Cuajimalpa	49.2793827160494
2	Magdalena Contreras	44.2960199004975
3	Tlalpan	38.5580491132333
4	Miguel Hidalgo	38.0464900662252
5	Cuauhtémoc	31.9897826086957
6	Tlalnepantla	30.3451914893617
7	Gustavo A. Madero	29.5074248927039
8	Ecatepec	25.1555042016807
9	Nezahualcóyotl	24.6968691588785
10	Milpa Alta	24.156835443038
11	Xochimilco	22.2418222222222
12	Texcoco	21.1502183406113

6. Documente el pseudocódigo que explique al menos
   - ⇒ Proceso con todos los pasos para realizar el proceso de integración de datos
   - ⇒ Reglas y validaciones usadas
   - ⇒ Modelo de datos empleado para catálogos y tabla de hechos principal
   - ⇒ Código fuente usado y capturas de pantallas que muestre la ejecución de los pasos principales
   - ⇒ Documente la exploración de lo datos realizados

Leer documentación de metadatos disponibles sobre el conjunto de datos a estudiar.

Crear catalogo de estaciones: estaciones.csv, con estructura: {id, Clave, Entidad, Delegación o Municipio, Estación

Descargar archivo REDDA.zip del año 2010 al 2019, posteriormente descomprimir cada archivo.

Obtener los archivos "*PPH.xls", se hace uso de la terminal de comandos para hacer dicha búsqueda y se crea la carpeta pph para el almacenamiento de los archivos requeridos.

Mkdir -p /Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph &
/Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset -name "*PPH.xls" -exec cp "{}"
/Users/aarongarcia/desktop/Data_Mining/practices/p4-etl/dataset/pph \;

Por medio de lenguaje de programación Python se realiza la integración de datos:

Por cada archivo *PPH.xls encontrado

Se abre el archivo

Eliminar espacios en blanco

Unificar tipos de fecha a YYYY-MM-DD HH:MM:SS

Extraer año de fecha y hacer nueva columna

Extraer mes de fecha y hacer nueva columna

Calcular numero de semana a partir de fecha y hacer nueva columna

Agregar dicha información a una lista global

Abrir archivo estaciones.csv

Crear un diccionario {clave de estación, identificador numérico}

Definir estructura de tabla de hechos: {"elemento", "fecha", "anio", "mes", "week", "localizacion", "medición"}

Combinar todos los elementos de la lista global

Por cada fila de la tabla obtenida al combinar la lista global:

Por cada estación de medición:

Agregar elemento (Para todos es Precipitación pluvial)

Agregar fecha (Misma para cada fila)

Agregar anio (Misma para cada fila)

Agregar mes (Misma para cada fila)

Agregar week (Misma para cada fila)

Agregar localización (Valor numérico del diccionario de acuerdo con la clave de estación)

Agregar medición (Valor numérico dado por la estación)

Para el caso de la exploración de datos, los queries se muestran en el punto #5, donde a manera de explicación del total de registros que se tienen es: 10 archivos con extensión .xls, 8 de ellos tienen 52 registros, da un sub total de 416 registros, 2 archivos tienen 53 registros, el segundo sub total es de 106 registros, sumamos estos dos subtotales y obtenemos 522 registros totales, dado que la tabla original de datos incluye a las estaciones como columnas, se tiene que hacer una transformación de únicamente estas 16 estaciones a columnas, por lo que cada uno de los 522 registros se multiplica por 16 estaciones, dando un total general de 8352 registros finales en la tabla de hechos.

Conclusiones

Muy tranquila la practica, cuando realice la exploración de datos para mi proyecto de AirBnB sin saberlo realice un ETL, ya que integre varios archivos y los transforme para poder analizarlos o explorarlos a mayor granularidad, considero que lo complicado es entender que es lo que quieres analizar o encontrar, ya con ello será como se organizara la tabla de hechos, los catálogos no siempre se pueden obtener de manera completa a partir de la tabla de hechos por lo que navegar en la fuente de datos será la mejor opción.