# **Proyecto 9:**

# **Extracción de Datos y Sentencias SQL vía Python**

## **Objetivos de aprendizaje y competencias**

* Obtención de datos a través de diversos métodos de extracción.
* Comprender que es una API y su funcionamiento.
* Generar queries de SQL en el entorno de Jupyter Notebook.
* Aprendizaje de sentencias JOIN.
* Conocer y manejar conceptos claves como request, JSON, cursors, fetchall.

## **Motivación**

El concepto **ETL** es un término estándar utilizado para referirse al movimiento y transformación de datos. Ello involucra el mover datos desde múltiples fuentes, reformatearlos y cargarlos en otra base de datos, con el objeto de analizarlos.

Como primer paso a la extracción de datos, se requiere conocer los diversos métodos para obtener estos. En este proyecto analizaremos tres formas de extracción las cuales son:

* Extracción de datos provenientes de una API (Interfaz de programación de aplicaciones).
* Extracción de datos provenientes de uno múltiples archivos locales (CSV’s, xlsx o otros formatos).
* Extracción de datos provenientes de una base de datos (En este caso, desde PostgreSQL).

Luego de la obtención de estos datos se deberán manipular a traves de sentencias de SQL utilizando la librería Psycopg2 para conectarnos con postgreSQL, donde a través de la interfaz Jupyter Notebook se realizará el enlace con la librería Pandas (Python).

## **Instrucciones**

**Parte 1 - Extracción de datos por medio de API**

Para este desafío obtendremos los datos que utilizaremos a traves de la API del sitio <https://earthquake.usgs.gov/>, que reporta información sobre sismos y terremotos a lo largo del mundo. Toda la documentación de la API en cuestión se encuentra en <https://earthquake.usgs.gov/fdsnws/event/1/>.

**1**. Obtenga los datos de todos los sismos ocurridos entre el 1 de Enero del 2010 y el día 5 de Diciembre del 2019, que tengan una magnitud mínima de 5 grados, en formato JSON (para efectos de la API, el formato a solicitar es el llamado “geojson”).mundo. Toda la documentación de la API en cuestión se encuentra en https://earthquake.usgs.gov/fdsnws/event/1/.

Para la obtención de los datos solicitados, se puede recurrir a la librería ***request****,* y para transformar aquella información en una variable de tipo diccionario, se puede utilizar la librería ***json****.* Lo anterior debería entregarle un diccionario con otros diccionarios al interior.

**2.** A partir de la información disponible en el diccionario generado, construya un DataFrame que muestre los registros de cada uno de los sismos antes mencionados.

Puede resultar conveniente obtener el número de registros extraídos disponible en el subdiccionario “metadata”, y recorrer el subdiccionario “features” con aquella información.

**3.** Genere con exactitud los siguientes dos gráficos para explorar el DataFrame indicado:

* Gráfico de barras horizontal que entregue los 25 lugares con mayor frecuencia de sismos en la década. Explore formas de cómo truncar la cantidad de datos a los 25 con mayor frecuencia.
* Histograma que muestre la distribución de sismos entre 6 grados y el máximo observado. Explore cómo fijar los márgenes del gráfico para que no se muestren en el eje x valores menores a 6 grados)

Genere además, un tercer gráfico que considere relevante para estos efectos, y explique las razones del porqué lo eligió.

**4.** Note que la columna “time” no está en formato “año/mes/dia hora/min/seg”. Para poder trabajar de mejor manera aquel dato, genere 6 nuevas columnas en el DataFrame, en la cual se guarden año, mes, dia, hora, min y seg de cada evento.

**5.** Construya dos nuevas columnas binarias:

* Una columna que tome valor 1 cuando el sismo ocurrió en Chile y tome valor 0 si no
* Una columna que tome valor 1 si el sismo tuvo una magnitud mayor que 6 grados y 0 si no.

**6.** Genere un nuevo DataFrame a partir del anterior, manteniendo solo las siguientes columnas:

* code
* mag
* alert
* place
* sig
* tsunami
* title
* las 6 columnas construidas en el punto 4.
* las 2 columnas binarias construidas en el punto 5.

**7.** Genere tres DataFrames donde cada uno almacene los sismos de los años 2010, 2015 y 2018 respectivamente. Expórtelos en 3 csv distintos. Nombrelos como “earthquakes\_mag6\_2010.csv”, “earthquakes\_mag6\_2015.csv” y “earthquakes\_mag6\_2018.csv” respectivamente.

**Parte 2 - Carga de los datos en SQL y ejecución de queries via Python**

Para el desarrollo de esta sección, seguiremos utilizando el ambiente de **Jupyter Notebook**. Para la extracción de los datos utilizaremos una base de datos la cual debe ser creada a través del motor PostgreSQL.

**8.** Generar una nueva base de datos con el siguiente nombre ‘earthquakes\_mag6’. Esto puede hacerlo mediante consola o a través de DBeaver.

**9.** Utilizando la librería psycopg2, genere una nueva conexión hacia el motor de PostgreSQL.

**10.** Utilizando la librería psycopg2, cree tres tablas con el nombre de ‘earthquakes\_mag6\_2010’’, ‘earthquakes\_mag6\_2015’ y ‘earthquakes\_mag6\_2018’’ e ingeste los datos con los csv exportados en el punto 6.en esta tabla. Revise la naturaleza de los datos ingestados utilizando dtypes.

**11.** A través de queries, realice las siguientes consultas:

* **Query 1:** En la tabla de sismos del año 2010, obtener los eventos igual o mayores a 7 grados donde se muestre lugar, magnitud, mes, dia. Ordénelos de manera descendente. Identifique el evento con mayor magnitud.
* **Query 2:** Identifique los sismos de igual magnitud que se hayan producido tanto en el año 2015 como en el año 2018.
* **Query 3:** Identifique los sismos que se han producido tanto en el año 2015 y 2018 y que tenga como alerta 'red'.
* **Query 4:** Identifique los sismos producidos el año 2018 con una magnitud sobre del 7 y riesgo de tsunami.

Los resultados de cada query deben mostrarlos como DataFrame.

**Parte 3 - Bonus**

**12. O**btenga nuevos datos de la API, considerando todos los sismos ocurridos entre el 1 de Enero del 1970 y el día 5 de Diciembre del 2019, que tengan una magnitud mínima de 5.5 grados. Restrinja además la solicitud de datos a aquellos eventos que se encuentren en longitudes entre -100 y -50 grados y en latitudes entre -55 y 20 grados.

A partir de ello, reproduzca el mismo gráfico de barras vertical, ahora con los 25 lugares con mayor frecuencia de sismos, y genere un histograma en función de los años que comprenden los datos, con un número de bins mayor o igual a 25. Analice los resultados y comente.

**Formato de entrega:**

* Entregar un archivo .ipynb con todo el código requerido para dar respuestas a las preguntas.
* Los 3 archivos csv que se solicitan en el enunciado.
* **Ojo:** No se deberá presentar luego de este proyecto. Por tanto, el **notebook debe ser autocontenido**, es decir, debe explicar de forma clara cada uno de los pasos de forma ordenada, contando con un relato claro que nos permita observar tanto el código, como los análisis y conclusiones de su trabajo.

**Evaluación:**

* Si bien no deberán presentar proyecto 8 ni 9, si tendrán una una evaluación luego del proyecto, con el objetivo de medir su capacidad de utilizar lo aprendido en el ambos proyectos para resolver preguntas similares con otras bases de datos, junto con verificar que comprenden los elementos técnicos más relevantes relacionados a la extracción de datos.

**Plazos de entrega:**

* Entrega Lunes 16/12/2019 hasta las 23:59.

**Links de referencia y de utilidad:**

[**https://earthquake.usgs.gov/earthquakes/feed/v1.0/geojson.php**](https://earthquake.usgs.gov/earthquakes/feed/v1.0/geojson.php)

[**https://es.python-requests.org/es/latest/**](https://es.python-requests.org/es/latest/)

[**https://www.dataquest.io/blog/loading-data-into-postgres/**](https://www.dataquest.io/blog/loading-data-into-postgres/)

[**https://programacionymas.com/blog/como-funciona-inner-left-right-full-join**](https://programacionymas.com/blog/como-funciona-inner-left-right-full-join)