**BITACORA DE EJECUCIÓN**

**TITULO**: WebScrapping y ETL Yellow Taxis

**OBJETIVO**: Descargar todos los archivos y generar el proceso de ETL a los data sets descargados en la web de NY.

<https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

**ARCHIVOS**:

* Webscrapping.ipynb → Archivo para descargar todos los data sets de yellow taxis, desde el 2019 al 2023.
* yellow\_tripdata\_2023-01.parquet → Archivo que se descarga de la web, para este proyecto descargamos del 2019 al 2023. Es un archivo por mes.
* DE\_yellowtaxis\_final\_ETL.ipynb: Primer codigo para generar el proceso de ETL a los archivos, este notebook tiene graficos y mas lectura de datos para comprender los mismos.
* DE\_yellowtaxis\_ETL\_GC.ipynb: Mismo código anterior, optimizado y sin graficos para ser trasladados a la función de la automatización.
* Funciones\_notebook.ipynb: Notebook donde se encuentran todas las funciones correspondientes al webscrapping, reducción de datos y proceso de ETL

**PROCEDIMIENTO**

**Parte Inicial: Primer descarga, lectura y conocimientos de datos**

1. Descargar los data sets de NY taxis amarillos.
2. Se generan la reducción del 5%
3. Se hace un merge para unir todos los datos
4. Luego se realizo el ETL, dejando el datasets final con los datos limpios.

**Segunda parte: Pasamos en limpio los ejecutables a funciones optimizadas para cargar los archivos en Google Storage**

1. Se optimiza el codigo de ETL para generar la funcion en el notebook de “DE\_yellowtaxis\_ETL\_GC.ipynb”
2. Se cargan las funciones con los códigos optimizados en el siguiente notebook “Funciones\_notebook.ipynb”
3. Ya con todas las funciones creadas y optimizadas en un notebook, se pasan al main.py para ser cargado en Google cloud y podes ser llamados a ejecutar en la máquina virtual creada en la plataforma.