

SEMANA 1 – GUÍA DE TRABAJO PRÁCTICO

Objetivos

- ✍ Conocer la infraestructura computacional del curso, basada en un ambiente Docker.
- ✍ Conocer diversos formatos de archivos, que permiten la definición de datos estructurados y semiestructurados.
- ✍ Realizar prácticas de procesamiento de datos grandes utilizando Python.

Prerrequisitos

- ✍ Conocimientos intermedios de programación en Python y habilidades de uso de notebooks. Esto es, uso de librerías, lectura y seguimiento de código ejemplo, desarrollo de código propio, depuración de programas, generación de resultados.
- ✍ Saber seguir rutas de localización de archivos y carpetas.
- ✍ Saber descargar documentos y subir entregables en la infraestructura de Coursera.

Metodología

- ✍ Se realiza de manera **individual**.
- ✍ Deben utilizarse la infraestructura, datos y *notebooks* entregados para el desarrollo de los retos propuestos.
- ✍ Las instrucciones de cada reto se encuentran descritas en el enunciado a continuación.
- ✍ No se reciben trabajos desarrollados en otras infraestructuras ni con otros conjuntos de datos.
- ✍ Se entrega en la infraestructura de Coursera, en el enlace previsto para tal fin.

Enunciado

Los trabajos prácticos del curso se realizan en una infraestructura virtual de GCP (Google Cloud Platform). Esto le permite trabajar en un entorno computacional que ya está configurado, de manera que su esfuerzo y tiempo se concentren en el desarrollo de los retos.

En el transcurrir del curso, todas las guías harán referencia al desarrollo, descarga y trabajo sobre dicha infraestructura. Por ello, es crítico que lo haga como primer paso para su trabajo práctico.

En la primera semana del curso se espera que usted desarrolle algunos retos sobre esta infraestructura. A continuación, encuentra la descripción de cada una de las actividades que debe lograr.

[1] Creación de la instancia en GCP

- ✍ Cree una instancia en la infraestructura de GCP. En la plataforma Coursera del curso, en la sección **Guías de Tecnología**, encuentra la guía “**GUÍA DE CREACIÓN DE INSTANCIA EN GCP**” para crear la instancia a partir de una imagen.
- ✍ Si tiene dificultades con el procedimiento, utilice Slack para resolver sus dudas.

[2] Procesamiento básico de datos

En este reto usted va a desarrollar ejercicios de trabajo básico de procesamiento de datos, con un *dataset* real y con algunas operaciones propuestas, que podrían corresponder a preguntas de negocio relacionadas con dichos datos.

En cada carpeta encuentra **tres conjuntos de datos**: el primero en formato .CSV, que corresponde a un archivo estructurado; el segundo en formato XML y el tercero en formato JSON, que corresponden a formatos semiestructurados.

Para cada conjunto de datos está propuesta y desarrollada **a manera de ejemplo** una pregunta de negocio, que se resuelve con el código Python que se le entrega. Enseguida encuentra un **reto propuesto como ejercicio** para su desarrollo. Usted debe desarrollar cada uno de esos ejercicios, de acuerdo con la guía particular que encuentra en el *notebook*. El resultado de la ejecución queda en una carpeta prevista dentro de su infraestructura y definida a continuación. Usted debe subir a Coursera la respuesta que se genera con su ejercicio, la cual será evaluada como parte de sus resultados en el curso.

En caso de tener dudas sobre los ejemplos, los *notebooks* o la manera de subir sus resultados a Coursera, utilice el canal del curso previsto en Slack.


Ubicación del notebook de trabajo: **Lo encuentra en la carpeta /home/ivanarturog620 (como se explica en la guía de creación de la instancia)**

Ruta/ubicación de resultados: **Una vez ejecuta el reto propuesto como ejercicio, dentro de la carpeta answer encontrará el archivo de respuesta .zip**




Entrega

Realice la entrega de los resultados utilizando el enlace previsto en Coursera **para cada uno** de los retos propuestos como ejercicio (procesamiento CSV, procesamiento JSON, procesamiento XML).

Manera de nombrar los archivos de resultados: **Answer.zip**

-  El archivo de respuesta debe seguir el nombre indicado para que se reconozca en el proceso automático de calificación. Asegúrese de subir a la plataforma de entrega en el enlace correcto cada uno de ellos.

Fecha y hora límite de entrega: **domingo** de la **semana 1** del curso a las **10 PM hora COLOMBIA**

-  Es la única forma válida de entrega.
-  Asegúrese de entregar cada uno de los resultados de los ejercicios en el enlace correspondiente. La calificación es realizada de forma automática. Si el nombre del archivo que se entrega o el contenido no es el esperado, su evaluación será 0/5.
-  No se reciben entregas tardías o incompletas a excepción de tener una excusa válida.