

SEMANA 2 – GUÍA DE TRABAJO PRÁCTICO

Objetivos

- ✍ Familiarizarse con herramientas como PySpark para el procesamiento y manipulación de datos que normalmente no caben en la memoria donde herramientas como Numpy y Pandas tienden a tener limitaciones.
- ✍ Realizar transformaciones y agregaciones con Spark.
- ✍ Almacenar el resultado final del ETL realizado.
- ✍ Realizar prácticas de procesamiento de datos grandes utilizando PySpark.

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 **Grupal**.
- ✍ 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 del Jupyter notebook de la semana 2.
- ✍ 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 la infraestructura de GCP, como se mencionó en la guía de la semana 1. 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.

[1] Iniciar actividad.

- ✍ Para esta actividad solo necesita en la carpeta s2 de su instancia.
- ✍ Para correr el contenedor corra el siguiente comando: `- docker-compose up`. En la terminal encontrará una URL de este estilo: `“http://<ip-de-su-maquina>:8888/lab?token=0f5dabdcdf9fd9941ade8ce267c9a61327c0ed4db82c1a0c”` Esta URL le da acceso al libro de jupyter para que pueda realizar la actividad. En caso de dudas referenciar la guía que encuentra en “Guía tecnologías” en Coursera bajo la sección “Arranque y logística”.

[1] ETL

¿Qué es un ETL?

Es un conjunto de herramientas y técnicas que nos permiten extraer (Extract) los datos de varias fuentes transformarlos (Transform) y cargarlos en una base de datos destino (Load), estas actividades son claves para obtener conocimiento de los datos, realizar clasificaciones, reformatear, filtrar, limpiar o entrenar modelos de inteligencia artificial.

En esta actividad va a recibir datos de **viajes en bicicleta, estaciones del clima** con información diaria del clima y **puntos de interés**, todo dentro de la ciudad de Nueva York, dentro del contenedor Docker encontrará más información como los diccionarios de datos. Se propone realizar una agregación de todos los datos mencionados para tener una matriz de datos más enriquecida y limpia, esta matriz de datos **posteriormente se va a utilizar para realizar unos entrenamientos de modelos para la semana 3**.

En *Jupyter notebook* apoyará en una parte del procesamiento de los datos (paso del 1 al 4 y pistas para el paso 5) sin embargo, usted debe terminar con la limpieza y **realizar** un informe de las decisiones tomadas por usted en el procesamiento y en la limpieza (por ejemplo, se eliminaron 1.000 filas ya que los datos eran nulos) y un ejemplo del resultado final de sus datos procesados (unas filas **ALEATORIAS**).

En caso de tener dudas sobre los ejemplos, los *notebooks* o la manera de subir sus resultados a Coursera, utilice Slack para atender sus dudas.

Ubicación del notebook de trabajo: `home/ivanarturo9620/s3`




Ruta/ubicación de los resultados del ETL: usted puede especificar la carpeta en el código, por ejemplo:
`"matriz_de_datos_final.write.parquet('output/end/travel_data_weather_station_end')"`

Entrega

Realice la entrega del PDF informe utilizando el enlace previsto en Coursera y entregue el Jupyter notebook con su código en un archivo zip.

Manera de nombrar los archivos de resultados: usuario correo uniandes. Por ejemplo, "jp.gonzales10_informe_s2.pdf" o "jp.gonzales10_la.rodriguez_informe_s4.pdf" para grupos.

Fecha y hora límite de entrega: **domingo** de la **semana 2** del curso a las **10 PM hora COLOMBIA**

-  Es la única forma válida de entrega.
-  Asegúrese de entregar el documento correspondiente.
-  No se reciben entregas tardías a excepción de tener una excusa valida.