Reto técnico

Instrucciones:

- Al recibir el reto, tendrás tres días para resolverlo y entregarlo.
- Siéntete libre de proponer la solución que consideres más adecuada según los requerimientos del ejercicio.
- Cualquier duda o ayuda que requieras, comunícate con José, el Lead de Data, al 5544509009.

Ejercicio

El equipo de Ciencia de Datos se ha acercado a ti ya que necesitan incorporar información sobre clima dentro de sus modelos predictivos. Para ello, han encontrado un sitio web del cual se puede obtener información. Te piden implementar un *scraper* básico y procesar los datos de la siguiente manera:

1) Cada hora debes ejecutar un proceso o corrida para obtener datos de varias ciudades, por lo que debes enviar peticiones y extraer el contenido de los siguientes enlaces:

    - https://www.meteored.mx/ciudad-de-mexico/historico
    - https://www.meteored.mx/monterrey/historico
    - https://www.meteored.mx/merida/historico
    - https://www.meteored.mx/wakanda/historico

    Para cada una de las peticiones que respondan con código 200 deberás extraer cuatro datos: distancia a la estación meteorológica, fecha y hora de actualización; temperatura actual y humedad relativa

Utiliza al menos una expresión regular para alguno de los datos requeridos, puedes utilizarla directamente sobre el texto, combinada con selectores HTML/CSS o algún otro método que creas conveniente.

Cada respuesta (exitosa o no) deberá ser guardada en un archivo JSON que incluya la petición que se realizó, el código de respuesta HTTP, los datos solicitados y algún identificador de la corrida (por ejemplo, 20230401_120000), en este punto no es necesario que los datos estén limpios.

2) Una vez generados los archivos JSON de la corrida, continúa con un proceso que estandarice los datos y los inserte en una base de datos SQL, el diseño debe considerar lo siguiente:

    - Un catálogo con las ciudades obtenidas y un identificador.
    - Una tabla en donde se guarden los códigos de las respuestas HTTP de cada una de las peticiones.
    - Una tabla con los datos obtenidos de las peticiones exitosas

3) Genera un archivo parquet con las siguientes especificaciones:

    - El archivo debe estar particionado de acuerdo a la corrida del proceso
    - Cada partición deberá contener un resumen de la información disponible en la base de datos al momento de la ejecución, concretamente, un registro por ciudad con la temperatura y humedad máximos, mínimos y promedio, así como última actualización de acuerdo al sitio

Requerimientos mínimos del ejercicio:

- Enlace al repositorio que contenga los recursos que resuelven el reto.
- Buenas prácticas en el manejo de excepciones, formas normales SQL, código modular y organización de datos.
- Automatizar el flujo de trabajo mediante alguna herramienta como Airflow.
- Archivo README o PDF donde se proporcione evidencia (diagramas, capturas de pantalla, comandos en consola, etc.) y se explique detalladamente el razonamiento de la solución, así como posibles mejoras y herramientas que considerarías para escalar, organizar y llevar tu solución al siguiente nivel.
- En una carpeta exclusiva anexar archivos CSV con los datos generados en la base de datos y parquet.

Requerimientos adicionales que puedes considerar:

- Empaquetar la solución con Docker o Docker Compose
- Agregar un proceso que ejecute pruebas de integridad o validaciones de los datos procesados

- Visualizar los datos mediante un tablero