CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Data Engineering

Módulo Nº1

Extracción y almacenamiento de datos

Objetivos

- Implementar **técnicas** de **extracción** de datos por medio del lenguaje de programación **Python**.
- Implementar **técnicas** de **almacenamiento** de datos, con el formato Delta lake.

Consigna

Desarrollar un programa en Python que realice:

1. **extracción** de una **API,** como fuente de datos,
2. convierta los datos obtenidos como DataFrames de Pandas
3. y los guarde de forma **cruda**, sin transformaciones o con leves transformaciones, en formato Delta lake.

Deberás usar la librería requests para obtener datos de **2 o más endpoints** de la misma API.

Uno de los endpoints debe devolver datos temporales, que se actualicen periódicamente (mínimo una vez al día), como por ejemplo: valores meteorológicos, cotizaciones de monedas o acciones de compañías, variaciones de índices económicos, estadísticas deportivas, etc, **El otro endpoint debe ofrecer datos estáticos** o metadatos, como por ejemplo campos o atributos que describen a una estación meteorológica, a una ciudad, a una empresa o moneda, a un club deportivo, etc..

Deberás realizar una extracción incremental y una full, según corresponda.

Además tendrás que guardar cada DataFrame en formato Delta lake, cada uno en un directorio específico, como si fuese que estás trabajando en un **data lake**.

- En caso de que estés haciendo una extracción incremental, deberás particionar por cada fecha y también por hora (si corresponde).

- En el caso de datos relativamente estáticos podes particionar, o no, por algún otro campo, si consideras necesario.

Podés elegir la API que quieras, siempre y cuando respete la consigna. En caso de tener dificultades para buscar alguna, podés pedir que se te asigne una, o bien solicitar el listado de APIs seleccionadas por estudiantes de ediciones anteriores del curso

Tener en cuenta que en la próximas entrega, habrá que realizar limpieza y procesamiento de los datos extraídos

Formato de presentación:

- Jupyter notebook (archivo .ipynb) o archivo Python (.py)
- Renombrar el archivo con: nombre y apellido, y el número del trabajo (TP1). Por ejemplo: GuidoFranco_TP1,
- El programa será ejecutado por el tutor desde una plataforma como Google Colab.
- Al subir el programa en el campus, deberán incluir un comentario o mensaje con los argumentos de las decisiones que hayan tomado tanto en la selección de la fuente de datos como en el desarrollo del programa de extracción.

Criterios de evaluación

1. Calidad del código y presentación
    a. El código de extracción debe estar bien estructurado y seguir buenas prácticas de programación en Python.
    b. El funcionamiento del código debe estar documentado de forma clara y concisa.
2. Implementación del programa de extracción
    a. Justificar la técnica de extracción seleccionada.
    b. El programa debe extraer los datos de la fuente seleccionada utilizando la técnica elegida.
3. Almacenamiento en Delta lake
    a. El código debe asegurar de la existencia de los directorios donde guardará los datos, sino los debe crear automáticamente
    b. El código debe pisar los datos en el directorio, o bien insertar nuevos registros según corresponda.