Data Engineering Test
Solvex 2023

Ejercicio 1: Manipulación de datos con Pandas y conjunto de datos de COVID-19

Utiliza Pandas y el conjunto de datos público de COVID-19 proporcionado por la Universidad de Johns Hopkins para realizar las siguientes tareas:

a) Descarga los datos de COVID-19 en formato CSV o JSON desde la URL pública.
b) Carga los datos en un DataFrame de Pandas.
c) Calcula el promedio de casos confirmados por día en un país específico.
d) Encuentra los 10 países con la tasa de mortalidad más alta (número de muertes / número de casos confirmados) hasta la fecha.

Puedes encontrar el conjunto de datos de COVID-19 en la URL pública, como la ofrecida por la Universidad de Johns Hopkins:
https://github.com/CSSEGISandData/COVID-19

Ejercicio 2: Procesamiento de datos con Spark y conjunto de datos de vuelos

Utiliza Apache Spark y un conjunto de datos público de vuelos, como el conjunto de datos de vuelos de 2015 proporcionado por la Oficina de Estadísticas de Transporte de EE. UU., para realizar las siguientes tareas:

a) Descarga los datos de vuelos en formato CSV desde la URL pública.
b) Carga los datos en un DataFrame de Spark.
c) Calcula la cantidad promedio de retrasos en la llegada de vuelos en un aeropuerto específico.
d) Encuentra las 10 rutas de vuelo más populares (pares de aeropuertos) en términos de la cantidad de vuelos.

Puedes encontrar conjuntos de datos de vuelos en sitios web de datos abiertos, como el Portal de Datos Abiertos del Gobierno de EE. UU.:
https://www.transtats.bts.gov/DL_SelectFields.asp

Ejercicio 3: Integración de Pandas y Spark con datos de películas y críticas

Supongamos que tienes dos conjuntos de datos: uno en Pandas y otro en Spark. El conjunto de datos de Pandas es una tabla llamada "datos_peliculas" con información sobre películas:

```
ID,Título,Año
1,Película1,2020
2,Película2,2019
3,Película3,2021
4,Película4,2018
```

El conjunto de datos de Spark es un DataFrame llamado "criticas" con información sobre las críticas de películas:

```
PeliculaID,Critico,Puntuacion
1,Critico1,4.5
2,Critico2,3.8
3,Critico1,4.2
4,Critico3,4.7
```

Combina estos dos conjuntos de datos para obtener una tabla que muestre el título de la película, el año de lanzamiento y la puntuación promedio de las críticas. Asegúrate de utilizar tanto Pandas como Spark en el proceso de integración.

Ejercicio de Web Scraping con Requests:

Supongamos que deseas obtener el precio actual del Bitcoin (BTC) de un sitio web de criptomonedas. La información que necesitas se encuentra en la página de https://coinmarketcap.com/currencies/bitcoin/.

Tu tarea es escribir un script en Python que realice lo siguiente:

1. Realiza una solicitud GET a la URL https://coinmarketcap.com/currencies/bitcoin/ utilizando la biblioteca `requests`.

2. Analiza el contenido de la página web para extraer el precio actual del Bitcoin.
3. Imprime el precio en la consola.

A continuación, un esqueleto del código para empezar:

```python
import requests
from bs4 import BeautifulSoup

# URL de la página de CoinMarketCap
url = 'https://coinmarketcap.com/currencies/bitcoin/'

# Realiza una solicitud GET para obtener el contenido de la página
response = requests.get(url)

# Verifica si la solicitud fue exitosa (código de respuesta 200)
if response.status_code == 200:
    # Analiza el contenido de la página web con BeautifulSoup
    soup = BeautifulSoup(response.text, 'html.parser')

    # Encuentra el elemento que contiene el precio actual del Bitcoin
    # Pista: Inspecciona la página web para identificar el elemento adecuado

    # Extrae el precio del elemento y almacénalo en una variable

    # Imprime el precio en la consola

else:
    print(f'Error al hacer la solicitud. Código de respuesta: {response.status_code}')
```

La tarea es completar el código, identificar el elemento HTML que contiene el precio del Bitcoin y extraer ese valor. Luego, imprime el precio en la consola.

Teoría

¿Cuál de las siguientes plataformas de Microsoft es una solución de análisis de big data en la nube?

a) Azure SQL Database
b) Azure Synapse Analytics
c) Azure Data Factory
d) Azure Active Directory

En el contexto de Azure Data Factory, ¿cuál de las siguientes actividades se utiliza para transformar y limpiar datos en un flujo de trabajo?

a) HDInsight Spark

b) Azure Databricks
c) Data Flow
d) Azure Stream Analytics

¿Cuál de las siguientes opciones es una característica clave de Apache Spark que permite procesar datos en memoria para un rendimiento más rápido?

a) Apache Hadoop
b) Apache Flink
c) Spark Streaming
d) Resilient Distributed Dataset (RDD)

En el contexto de Pandas, ¿cuál de las siguientes operaciones se utiliza para eliminar filas duplicadas de un DataFrame?

a) df.groupby()
b) df.drop_duplicates()
c) df.fillna()
d) df.pivot_table()

¿Qué lenguaje de programación se utiliza comúnmente en Azure Databricks para el procesamiento de datos y análisis?

a) R
b) Java
c) Scala
d) C#