Практична робота №2

Вивчення основ агрегації даних за допомогою PySpark і візуалізація результатів в Jupyter Notebook.

Хід роботи

1. Повторити кроки запуску Jupyter Notebook з практичної роботи №1.

Переконатися, що дані з cities.csv доступні в папці.

Відкрити новий Jupyter notebook і зберегти його як data_aggregation.ipynb.

Завантажити дані з файлу cities.csv в DataFrame.

Вивести структуру DataFrame та перші кілька записів для перевірки.

2. Обчислити сумарне населення по регіонах.

Знайти місто з найбільшою площею в кожному регіоні.

Порахувати середню густоту населення по всіх містах.

```
from pyspark.sql import SparkSession
from pyspark.sql import functions as f
spark = SparkSession.builder.master("local").appName("PR1").getOrCreate()
citiDF = spark.read.format("csv") \
.option("header", "true") \
.option("inferSchema", "true") \
.load("C:/Users/WSuser/Desktop/vnau_edu_docs/inteligent_systems/practica/data/cities.csv")
citiDF.cache()
citiDF.printSchema()
citiDF.show(3)
citiDF.select("CityName", "Population").show()
citiDF.groupBy("Region").agg(F.sum("Population").alias("TotalPopulation")).show()
citiDF.sort("Population", ascending=False).show()
citiDF.filter(citiDF["IsCapital"] == True).show()
```

3.* Індивідуальна задача. Вибрати тему та згенерувати джерело даних по типу як у поточній практичній роботі у форматі *.csv та спробувати завантажити, задати структуру та вивести на екран:

- Автомобілі (модель, виробник, рік випуску, тип пального, ціна)
- Кліматичні дані (дата, температура, вологість, опади, швидкість вітру)
- Курси валют (дата, валюта, курс до USD, курс до EUR, зміна відсотків)
- Книги (назва, автор, рік видання, жанр, кількість сторінок)
- Фільми (назва, режисер, рік випуску, жанр, рейтинг IMDb)
- Технологічні продукти (назва, категорія, ціна, дата випуску, виробник)
- Ресторани (назва, місцезнаходження, кухня, середній чек, рейтинг)
- Університети (назва, країна, рейтинг, кількість студентів, рік заснування)
- Спортивні команди (назва, місто, вид спорту, кількість чемпіонатів, рік заснування)
- Музичні альбоми (назва, виконавець, жанр, рік випуску, тривалість)
- Туристичні пакети (назва, місцезнаходження, ціна, тривалість, включені послуги)
- Телефонні контакти (ім'я, номер телефону, електронна пошта, адреса, день народження)
- Лікарські засоби (назва, виробник, тип, вартість за одиницю, дата випуску)

- Погодні станції (назва, місцезнаходження, висота над рівнем моря, тип сенсорів, рік заснування)
- Електронні пристрої (модель, виробник, тип пристрою, ціна, рік випуску)
- Історичні події (дата, подія, місцезнаходження, учасники, наслідки)
- Аеропорти (назва, місто, країна, кількість рейсів на добу, рік заснування)
- Медичні заклади (назва, тип закладу, адреса, кількість лікарів, спеціалізація)
- Наукові дослідження (назва, галузь, рік початку, основні відкриття, керівник проекту)
- Політичні партії (назва, країна, лідер, рік заснування, ідеологія)

Висновки

Замість висновків, ви усі молодці!