МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний аерокосмічний університет ім. М.Є. Жуковського

“Харківський авіаційний інститут”

Кафедра комп’ютерних систем, мереж і кібербезпеки

Лабораторна робота № 2

з дисципліни “ Великі дані для кібербезпеки”

Варіант № \_2\_\_

|  |  |  |
| --- | --- | --- |
| Виконав студент | гр. 555іM | Орлов С.В. |
|  | (№ групи) | (П.І.Б.) |
| (підпис, дата) |  |  |
| Перевірив | д.т.н., професор, професор каф. 503 | |
|  | (науковий ступінь, вчене звання, посада) | |
|  | Фесенко Г. В. | |
| (підпис, дата) | (П.І.Б.) | |

2023

**Тема роботи**

Створення та робота з DataFrames з використанням Spark.

**Мета роботи**

Отримати навички створення та роботи з DataFrames з використанням Spark.

**Завдання**

Варіант 12 = 2

Перед виконанням роботи було завантажено та встановлено Spark на ОС Linux.

1.2 Створення DataFrame

1.2.1 Створення DataFrames з JSON – файлів

Зображення, що містить текст, знімок екрана, програмне забезпечення

Автоматично згенерований опис

Зображення, що містить текст, знімок екрана, програмне забезпечення, Мультимедійне програмне забезпечення

Автоматично згенерований опис

1.2.2.1 Створюємо DataFrame список, що включає сім назв місяців року англійською мовою, з яких не менше двох назв однакової довжини

Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис

1.2.2.2 Перетворимо отриманий список в RDD, і потім застосуємо

функцію map для отриманого RDD, щоб отримати новий RDD, який містить

пари «колір-довжина»:

Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис

1.2.2.3 Показуємо отриманий результат:

Зображення, що містить текст, знімок екрана, Шрифт, число

Автоматично згенерований опис

1.2.2.4 Здійснемо підрахунок кількості рядків в отриманому DataFrame:

Зображення, що містить текст, Шрифт, знімок екрана, чорний

Автоматично згенерований опис

1.2.2.5 Створимо масив з назв стовпців:

Зображення, що містить текст, знімок екрана, Шрифт, чорний

Автоматично згенерований опис

1.2.2.6 Вилучимо стовпець з назвою "length":

Зображення, що містить текст, знімок екрана

Автоматично згенерований опис

1.2.2.7 Переведемо формат DataFrame в JSON-формат



1.2.2.8 Залишимо тільки рядки стовпця "month" з довжиною 4 або 5, виберемо стовпець " month " і перейменуємо його в "mid\_length" і покажемо результат:

Зображення, що містить текст, знімок екрана

Автоматично згенерований опис

1.2.2.9 Вилучимо рядок зі словом "March" і покажемо результат:

Зображення, що містить текст, програмне забезпечення, Мультимедійне програмне забезпечення, знімок екрана

Автоматично згенерований опис

1.2.2.10 Відсортуємо за алфавітом назви місяців:

Зображення, що містить текст, знімок екрана, Шрифт

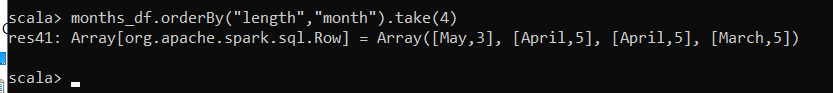
Автоматично згенерований опис

1.2.2.11 Залишимо назви місяців з кількістю літер у назві >=4 і відсортуємо назви місяців в порядку зростання довжини їх назви:

Зображення, що містить текст, знімок екрана

Автоматично згенерований опис

1.2.2.12 Об’єднаємо в пари назви місяців і довжину їх назви і виберемо чотири такі пари:



1.2.2.13 Відсортуємо місяці в порядку зменшення довжини їх назви:

Зображення, що містить текст, знімок екрана, програмне забезпечення, Мультимедійне програмне забезпечення

Автоматично згенерований опис

1.2.2.14 Визначимо скільки існує слів для кожної довжини слова:

Зображення, що містить текст, знімок екрана

Автоматично згенерований опис

1.3 Застосування DataFrame для аналізу даних

1.3.1 Підготуємо три DataFrame з RDD і покажемо, що вийшло для кожного з них.

Зображення, що містить текст, знімок екрана, програмне забезпечення, Мультимедійне програмне забезпечення

Автоматично згенерований опис

Зображення, що містить текст, знімок екрана, програмне забезпечення

Автоматично згенерований опис

Зображення, що містить текст, знімок екрана

Автоматично згенерований опис

1.3.2 Об'єднаємо дані трьох таблиць в одну:

Зображення, що містить текст, знімок екрана, програмне забезпечення, Мультимедійне програмне забезпечення

Автоматично згенерований опис

final\_data.sort("emp\_id").show()

Зображення, що містить текст, знімок екрана, програмне забезпечення, Мультимедійне програмне забезпечення

Автоматично згенерований опис

1.3.3 Створимо новий Data Frame, що містить відсутні значення для певних параметрів, для чого:

1.3.3.1 Визначимо case class Author:

Зображення, що містить текст, знімок екрана, Шрифт

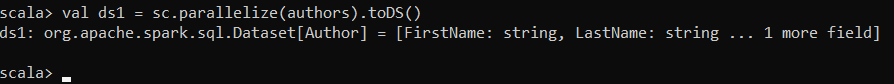
Автоматично згенерований опис

1.3.3.2 Задамо масив даних:

Зображення, що містить текст, Шрифт, знімок екрана

Автоматично згенерований опис

1.3.3.3 Створимо Data Frame і потім покажемо результат



Зображення, що містить текст, знімок екрана, програмне забезпечення, Шрифт

Автоматично згенерований опис

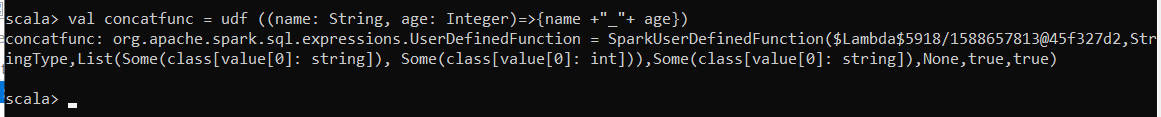
1.3.4 Приберемо в ньому рядки, де відсутнє хоча б одне значення:

Зображення, що містить знімок екрана, текст

Автоматично згенерований опис

1.3.5. Проведемо об'єднання стовпців Data Frame з 1.3.2, для чого:

1.3.5.1 Використовуючи функцію udf, покажемо як будуть представлені об'єднані дані:



1.3.5.2 Проведемо об'єднання стовпців "name"і "age", вказавши нову назву для об’єднаного стовпця:

Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис

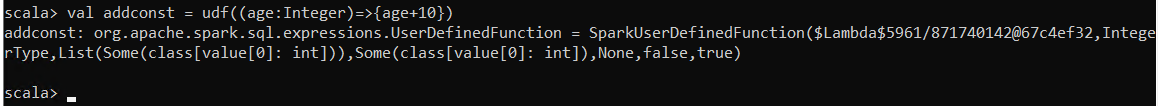
1.3.5.3 Покажемо, що вийшло:

Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис

1.3.6 Додамо константу до віку ("age") і запишемо отримане значення в новому стовпці, для чого:

1.3.6.1 Використовуючи функцію udf, покажемо як буде відбуватися збільшення віку:



1.3.6.2 Виконаємо додавання віку і помістимо дані в новий стовпець:

Зображення, що містить текст, знімок екрана, Шрифт, ряд

Автоматично згенерований опис

1.3.6.3 Покажемо, що вийшло:

Зображення, що містить текст, знімок екрана

Автоматично згенерований опис

1.3.7 Відсортуємо список в порядку зростання віку співробітників:

Зображення, що містить текст, знімок екрана

Автоматично згенерований опис

1.3.8 Представимо список співробітників в алфавітному порядку:

Зображення, що містить текст, знімок екрана, Шрифт, програмне забезпечення

Автоматично згенерований опис

1.3.9 Відсортуємо список в порядку зростання заробітної плати співробітників:

Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис

1.3.10 Додамо новий стовпець "Initial" і відсортуємо в алфавітному порядку ініціали співробітників:

Зображення, що містить текст, знімок екрана, програмне забезпечення, Мультимедійне програмне забезпечення

Автоматично згенерований опис

1.3.11 Визначимо середній вік і середню зарплату працівників:

Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис

1.3.12 Визначимо суму витрат на зарплату співробітників:

Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис

1.3.13 Визначимо кількість співробітників на кожній з посад:

Зображення, що містить текст, знімок екрана, Шрифт, число

Автоматично згенерований опис

1.3.14 Залишимо в списку тільки тих співробітників, у кого зарплата не менше 12000:

Зображення, що містить текст, знімок екрана, програмне забезпечення, Шрифт

Автоматично згенерований опис

1.3.15 Залишимо в списку тільки тих співробітників, у кого вік від 25 до 32 років включно та відсортуємо їх в порядку збільшення віку:

Зображення, що містить текст, знімок екрана, Шрифт, програмне забезпечення

Автоматично згенерований опис

**Висновки**

Було розглянуте поняття розподілених наборів даних, процес установки Apache Spark на ОС Лінукс, синтаксис Scala. Отримані знання та навички роботи з DataFrames. Відпрацьовані завдання за прикладами.