МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний аерокосмічний університет ім. М.Є. Жуковського

“Харківський авіаційний інститут”

Кафедра комп’ютерних систем, мереж і кібербезпеки

Лабораторна робота № 1

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**

Програмування операцій зі стійкими розподіленими наборами даних (Resilient Distributed Datasets, RDD) з використанням Spark.

**Мета роботи**

Отримати навички програмування операцій з RDD, а також навчитися працювати з парами «ключ-значення».

**Завдання**

Варіант 12 = 2

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

1.2.1.1 Створення RDD самостійно без вказівки рівня паралелізму:

*Виконайте контрольний приклад, після чого створіть свій RDD з 7-ми елементів трьома можливими способами.*

|  |
| --- |
| *Зображення, що містить текст, знімок екрана, Шрифт  Автоматично згенерований опис* |

1.2.1.2 Створення RDD самостійно із зазначенням рівня паралелізму:

*Виконайте контрольний приклад, після чого створіть свій RDD з 8-ми елементів і рівнем паралелізму 4 трьома можливими способами.*

|  |
| --- |
|  |

1.2.1.3 Створення RDD шляхом завантаження текстового файлу:

*Виконайте контрольний приклад, після чого завантажте цей файл, вказавши свій власний шлях до нього.*

|  |
| --- |
|  |

1.2.2.1 Функції filter, collect, first

*Виконайте контрольний приклад, після чого: створіть свій RDD, який складається з елементів від 2 до 20 з кроком 3; відфільтруйте отриманий RDD за умовою «всі елементи, що діляться на 2»; поверніть всі елементи нового масиву.*

|  |
| --- |
|  |

*Виконайте контрольний приклад, після чого виконайте фільтрацію по слову "scala", вибравши потім перше слово з профільтрованого RDD.*

|  |
| --- |
|  |

1.2.2.2 Функція distinct

*Виконайте контрольний приклад, а потім створіть набір RDD (з не менше ніж 5 елементами і не менше ніж одним повтором назви) згідно варіанту і виконайте дії, аналогічні контрольному прикладу (необхідно використовувати назви, що складаються тільки з одного слова).*

Варіант 12 = 2: з назв штатів США англійською мовою.

|  |
| --- |
|  |

1.2.2.3 Функція intersection для двох наборів

*Виконайте контрольний приклад, після чого створіть свої RDD (перший з числами від 8 до 16 з кроком 1, другий – з числами від 5 до 13 з кроком 2) і отримаєте новий набір RDD, що містить тільки елементи, присутні в обох вихідних наборах.*

|  |
| --- |
|  |

1.2.2.4 Функція union для двох наборів

*Виконайте контрольний приклад, після чого створіть свої RDD і виконайте їх об'єднання, використовуючи два способи, розглянутих вище.*

|  |
| --- |
|  |

1.2.2.5 Функція cartesian для двох наборів

*Створіть свої RDD (з трьома значеннями у кожному) і виконайте дії, аналогічні контрольного прикладу.*

|  |
| --- |
|  |

1.2.2.6 Перетворення map і flatMap

*Виконайте контрольний приклад, після чого створіть свій RDD і проведіть для нього обчислення квадратів.*

|  |
| --- |
|  |

*Створіть свої RDD, аналогічні розглянутим і виконайте для них дії трьох прикладів, розглянутих вище.*

|  |
| --- |
|  |

1.2.3 Агрегування у Spark

Таблиця 1 – завдання до варіанту 3

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

Автоматично згенерований опис

|  |
| --- |
|  |

1.2.4.2 Створення в Scala набору пар з рядків, де роль ключа відіграє перше слово:

*Виконайте контрольний приклад, а потім створіть набір RDD (з не менше ніж 5 елементів) згідно варіанту і виконайте дії, аналогічні контрольному прикладу (необхідно використовувати назви, що складаються тільки з одного слова):*

Варіант 1: з назв штатів США англійською мовою.

|  |
| --- |
|  |

1.2.4.4 Групування значень з однаковими ключами за допомогою groupByKey для набору пар (в ролі ключа виступає довжина слова)

*Створіть набір RDD (з не менше 5 елементів) згідно варіанту і виконайте groupByKey, countByKey, collectAsMap, lookup(key) (прийміть його рівним 3) (необхідно використовувати назви, що складаються тільки з одного слова):*

|  |
| --- |
|  |

1.2.4.5 Отримання набору з одними ключами за допомогою keys

*Виконайте контрольний приклад, а потім створіть набір RDD (з не менше ніж 5 елементів) згідно варіанту і виконайте дії, аналогічні контрольному прикладу (необхідно використовувати назви, що складаються тільки з одного слова):*

Варіант 2: з назв днів тижня англійською мовою

|  |
| --- |
|  |

1.2.4.6 Виконання внутрішнього з’єднання двох наборів пар за допомогою join

*Створіть два набори RDD (з не менше ніж 3 елементів у кожному з них) згідно варіанту і виконайте дії, аналогічні контрольному прикладу (необхідно використовувати назви, що складаються тільки з одного слова):*

Варіант 2: перший набір – з назв планет Сонячної системи англійською мовою; другий набір – з назв овочів англійською мовою.

|  |
| --- |
|  |

1.2.4.7 Виконання з'єднання двох наборів RDD, де ключ повинен бути представлений в першому RDD за допомогою leftOuterJoin

|  |
| --- |
|  |

1.2.4.8 Виконання з'єднання двох наборів RDD, де ключ повинен бути представлений в першому RDD за допомогою rightOuterJoin

|  |
| --- |
|  |

**Висновки**

Було розглянуте поняття розподілених наборів даних, процес установки Apache Spark на ОС Лінукс, синтаксис Scala. Відпрацьовані завдання за прикладами.