

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет інформатики та обчислювальної техніки
Кафедра інформатики та програмної інженерії

Лабораторна робота №4

з дисципліни: «Технології паралельного програмування в умовах великих даних»

з теми: «Big Data з використанням засобів Apache Spark»

Перевірив:
доцент
Жереб К.А.

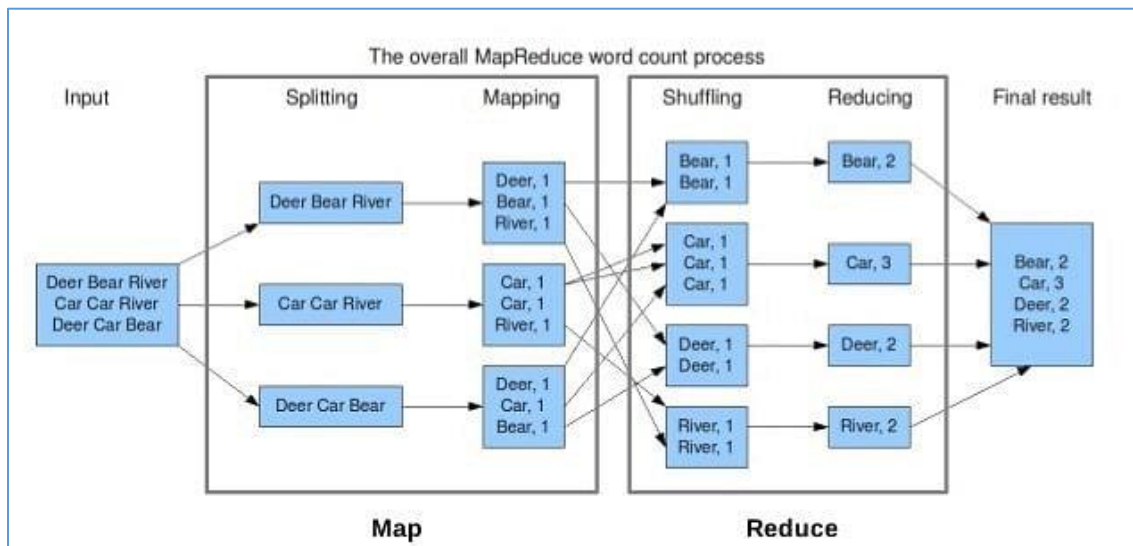
Виконав:
студент групи IT-01мн
Корзун І.М.

Завдання

- Обрати задачу та реалізувати для неї рішення з використанням технології Apache Spark

Хід роботи

В якості задачі обрано підрахунок слів у текстах. В умовах послідовного виконання дану проблему можна оцінити квадратичною складністю, бо для кожного слова треба порахувати кількість повторень у вхідному тексті. В якості оптимізації часу використаємо технології Apache Spark, яка аналогічно попередній роботі, здатна помітно скоротити час виконання.



- Рисунок 1. MapReduce.

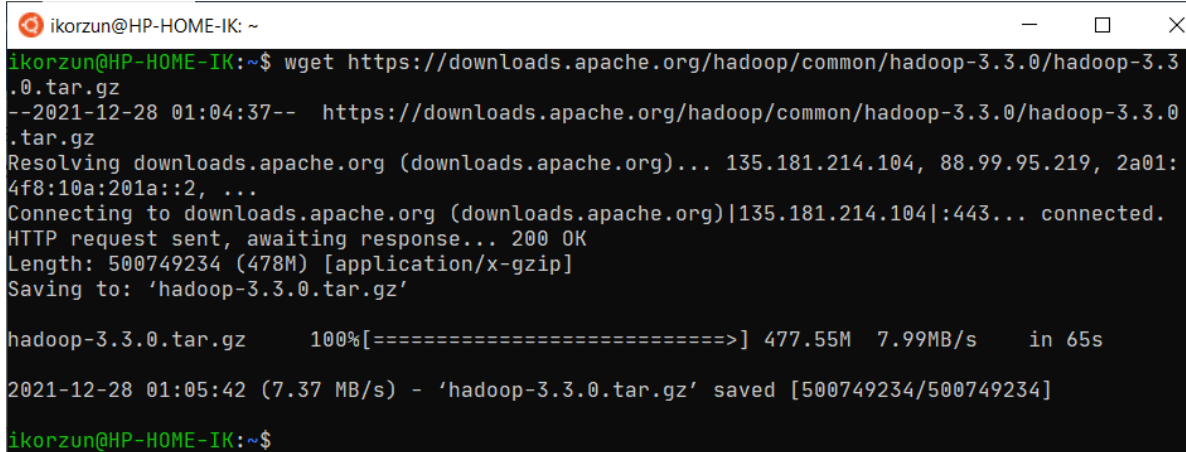
При виконанні поставленого завдання, спираючись на результати попередньої роботи, було вирішено наступні задачі:

- встановлення і налаштування Apache Spark:
 - завантаження останньої версії [рушія з офіційної веб-сторінки](#)
 - визначення системних змінних
- встановлення бібліотеки PySpark для реалізації задачі мовою Python
- реалізація задачі
- тестування програми

Встановлення і налаштування Apache Spark

- завантаження актуальної збірки Apache Spark

```
$ wget https://d1cdn.apache.org/spark/spark-3.2.0/spark-3.2.0-bin-hadoop3.2.tgz
```

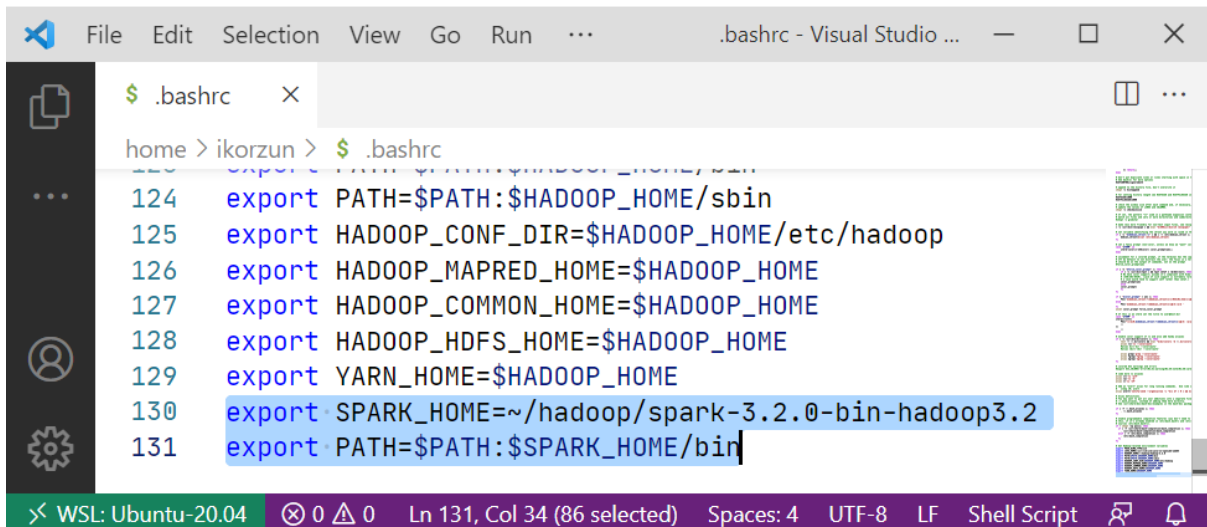


A terminal window titled 'ikorzun@HP-HOME-IK: ~' showing the execution of the command `wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz`. The output shows the file being resolved, connected to, and downloaded successfully. The progress bar indicates 100% completion with a speed of 7.99 MB/s. The file size is 500749234 bytes (478M).

```
$ tar -xvzf spark-3.2.0-bin-hadoop3.2.tgz -C ~/hadoop
```

- додавання системних змінних, які посилатимуться на директорію збірки

```
$ code ~/.bashrc
```



A screenshot of the Visual Studio Code editor showing the `~/.bashrc` file. The file contains several `export` statements for Hadoop and Spark. The line `export SPARK_HOME=~/.hadoop/spark-3.2.0-bin-hadoop3.2` is highlighted. The status bar at the bottom indicates the file is in the `WSL: Ubuntu-20.04` environment.

```
$ source ~/.bashrc
```

Налаштування інтерпретатора Python

- встановлення бібліотеки PySpark

```
$ python3 -m pip install pyspark
```

```
ikorzun@HP-HOME-IK:~$ python3 -m pip install pyspark
Requirement already satisfied: pyspark in ~/.local/lib/python3.8/site-packages (3.2.0)
Requirement already satisfied: py4j==0.10.9.2 in ~/.local/lib/python3.8/site-packages (from p
yspark) (0.10.9.2)
ikorzun@HP-HOME-IK:~$
```

Реалізація обраної задачі за допомогою бібліотеки PySpark мовою Python

```
import sys
from operator import add

from pyspark.sql import SparkSession

if __name__ == "__main__":
    if len(sys.argv) != 2:
        print("Usage: wordcount <file>", file=sys.stderr)
        sys.exit(-1)

    spark = SparkSession\
        .builder\
        .appName("PythonWordCount")\
        .getOrCreate()

    lines = spark.read.text(sys.argv[1]).rdd.map(lambda r: r[0])
    counts = lines.flatMap(lambda x: x.split(' ')) \
        .map(lambda x: (x, 1)) \
        .reduceByKey(add) \
        .sortBy(lambda a: -a[1]) \

    output = counts.take(10)
    for (word, count) in output:
        print("%s: %i" % (word, count))

    spark.stop()
```

- створення директорії вхідних даних у файловій системі Hadoop (hdfs)

```
$HADOOP_HOME/bin/hdfs dfs -mkdir input  
$HADOOP_HOME/bin/hdfs dfs -put $lab/book.txt input
```

- тестування програми (на відміну від попередньої роботи, можливе тільки у платформі Apache Hadoop)

```
ikorzun@HP-HOME-IK:~$ python3 $lab/wordcount.py /user/ikorzun/input/book.txt  
2021-12-28 04:12:36,103 WARN util.Utils: Your hostname, HP-HOME-IK resolves to a loopback add  
ress: 127.0.1.1; using 172.30.22.61 instead (on interface eth0)  
2021-12-28 04:12:36,104 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another ad  
dress  
WARNING: An illegal reflective access operation has occurred  
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/home/ikorzun/ha  
doop/spark-3.2.0-bin-hadoop3.2/jars/spark-unsafe_2.12-3.2.0.jar) to constructor java.nio.Dire  
ctByteBuffer(long,int)  
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform  
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access op  
erations  
WARNING: All illegal access operations will be denied in a future release  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
2021-12-28 04:12:37,748 WARN util.NativeCodeLoader: Unable to load native-hadoop library for  
your platform... using builtin-java classes where applicable  
: 14639  
the: 7906  
of: 5425  
and: 2759  
a: 2422  
to: 2168  
is: 2068  
in: 2048  
that: 1273  
are: 921
```

- порівняння з попередньою роботою за часом виконання

```
python3 $lab3/mostusedword.py -r hadoop $lab/book.txt  
83.47220635414124 msecs  
  
python3 $lab4/wordcount.py /user/ikorzun/input/book.txt  
8.051442623138428 msecs
```

Висновок

У результаті виконання лабораторної роботи реалізовано рішення для підрахунку слів з використанням підходу MapReduce на платформі Apache Hadoop за допомогою технології Apache Spark. У ході роботи вивчено особливості встановлення, налаштування та застосування технології та окремо методи

створення рішень із залучення Hadoop до обчислень на великих даних за допомогою мови Python, хоча для даного інструменту Scala є більш нативним методом розробки.