

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

Лабораторна робота №3

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

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

Перевірив:
доцент
Жереб К.А.

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

Завдання

- Обрати задачу та реалізувати для неї рішення з використанням підходу MapReduce та технології Apache Hadoop

Хід роботи

В якості задачі обрано підрахунок слів у текстах. В умовах послідовного виконання дану проблему можна оцінити квадратичною складністю, бо для кожного слова необхідно визначити кількість повторень у вхідному тексті. У якості оптимізації часу використаємо технології Apache Hadoop, яка за допомогою методів, апробованих у двох попередніх лабораторних роботах, здатна помітно скоротити час виконання.

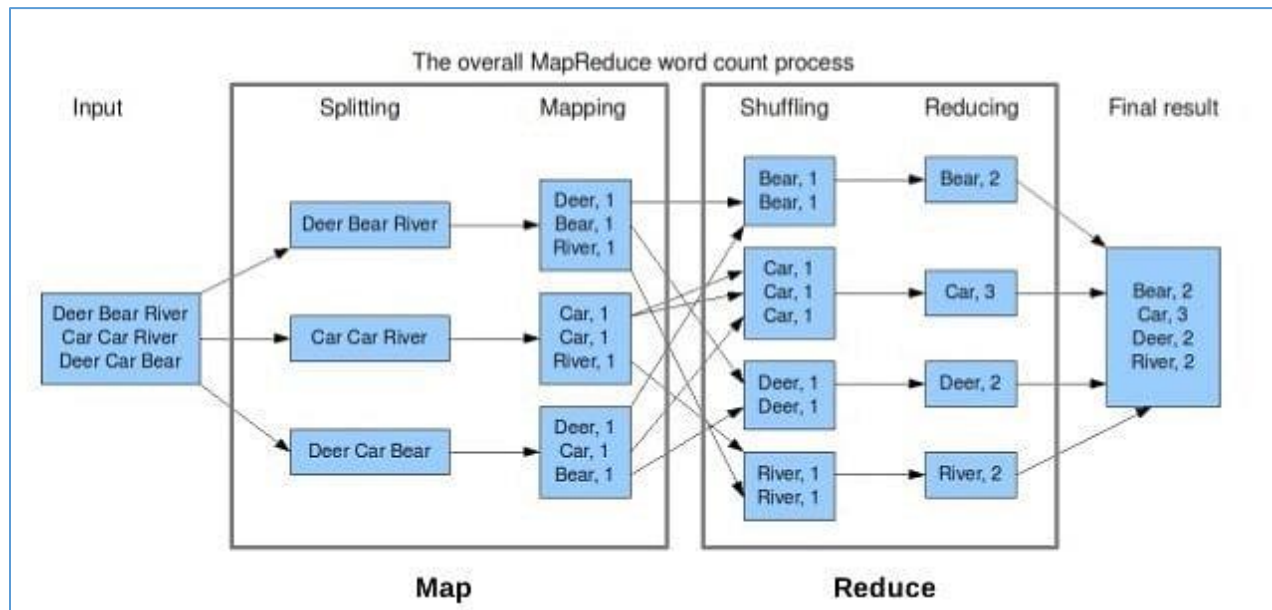


Рисунок 1. Модель розподілення роботи MapReduce, згідно якої працює платформа Hadoop

При виконанні поставленого завдання було вирішено наступні задачі:

- налаштування робочого середовища
 - встановлення прошарку сумісності в операційній системі Windows для запуску виконуваних файлів Linux (WSL, Windows subsystem for Linux)
- офіційна версія Hadoop не містить виконуваних файлів для Windows, через що їх необхідно збирати окремо – навпроти, в Linux вони вже прекомпільовані*

- встановлення і налаштування Apache Hadoop:
 - встановлення необхідних модулів
 - підтримуваної версії Java
 - засобу організації віддаленого доступу OpenSSH для організації взаємодії між сервісами Hadoop та клієнтом
 - завантаження актуальної версії платформи Apache Hadoop
 - розміщення збірки в підсистемі Linux (Ubuntu)
 - конфігурація та запуск
 - встановлення інтерпретатора Python (версії 3.x)
 - встановлення бібліотеки MrJob, яка надає зручний інтерфейс для реалізації моделі MapReduce
 - реалізація моделі MapReduce для обраної задачі
 - тестування програми

Налаштування робочого середовища

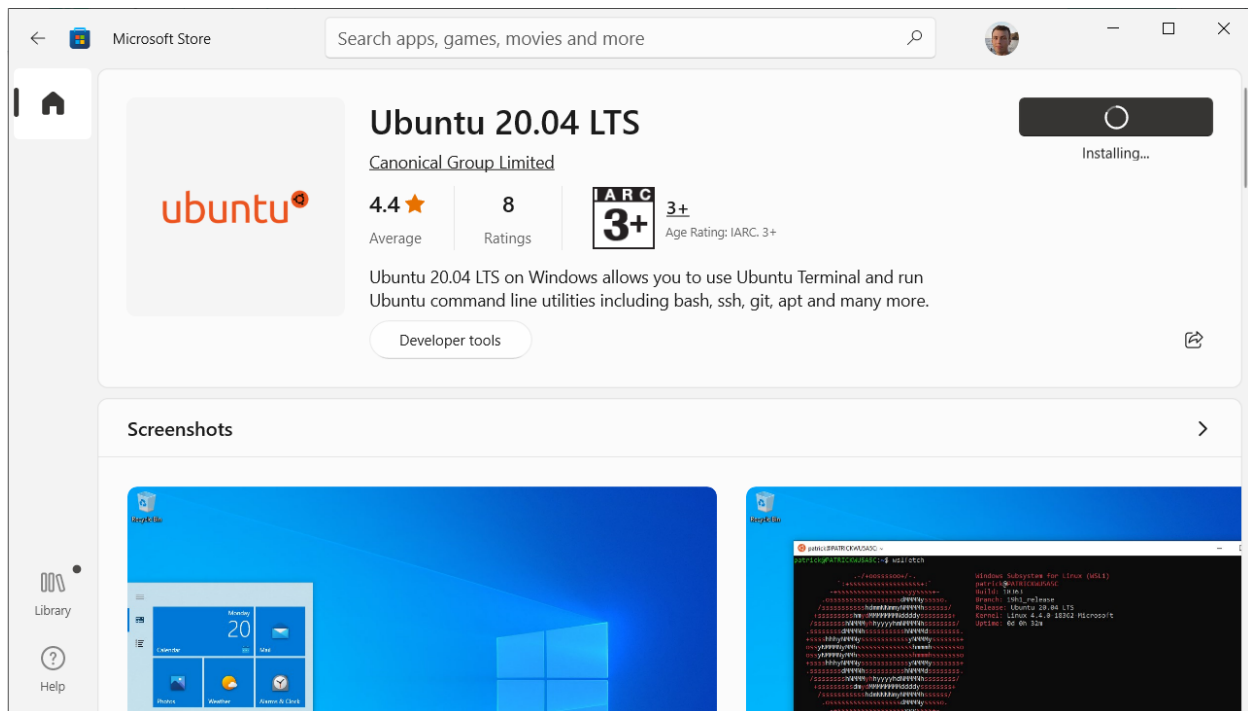


Рисунок 2. Завантаження підсистеми Linux (Ubuntu) Windows 10 через додаток Microsoft Store.

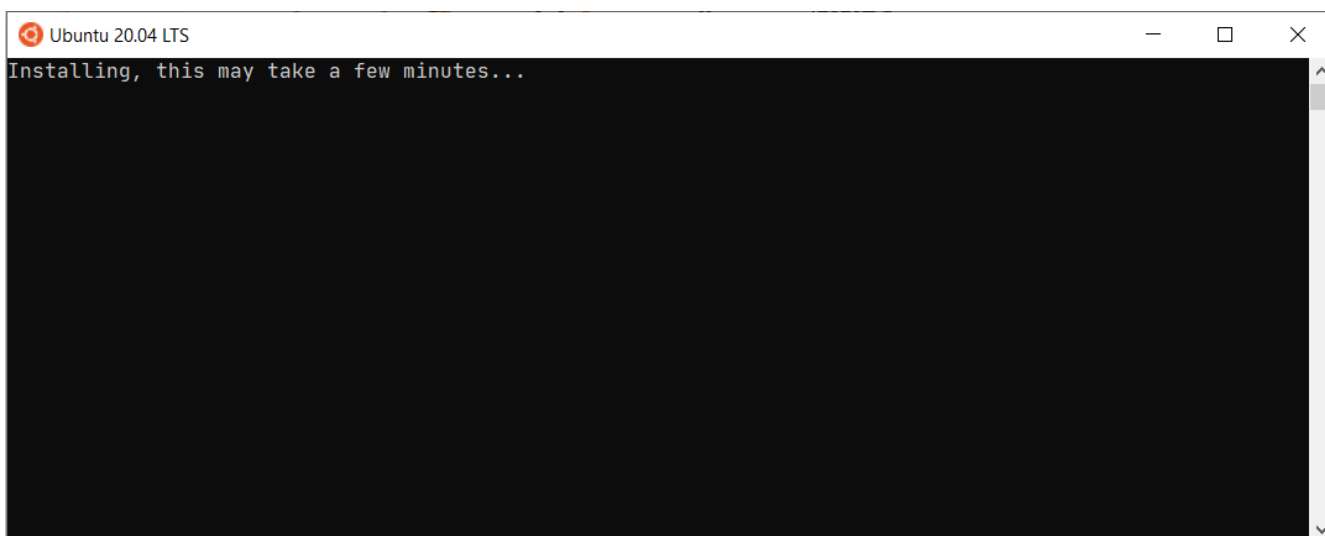


Рисунок 3. Встановлення операційної системи Ubuntu у WSL при першому запуску програми.

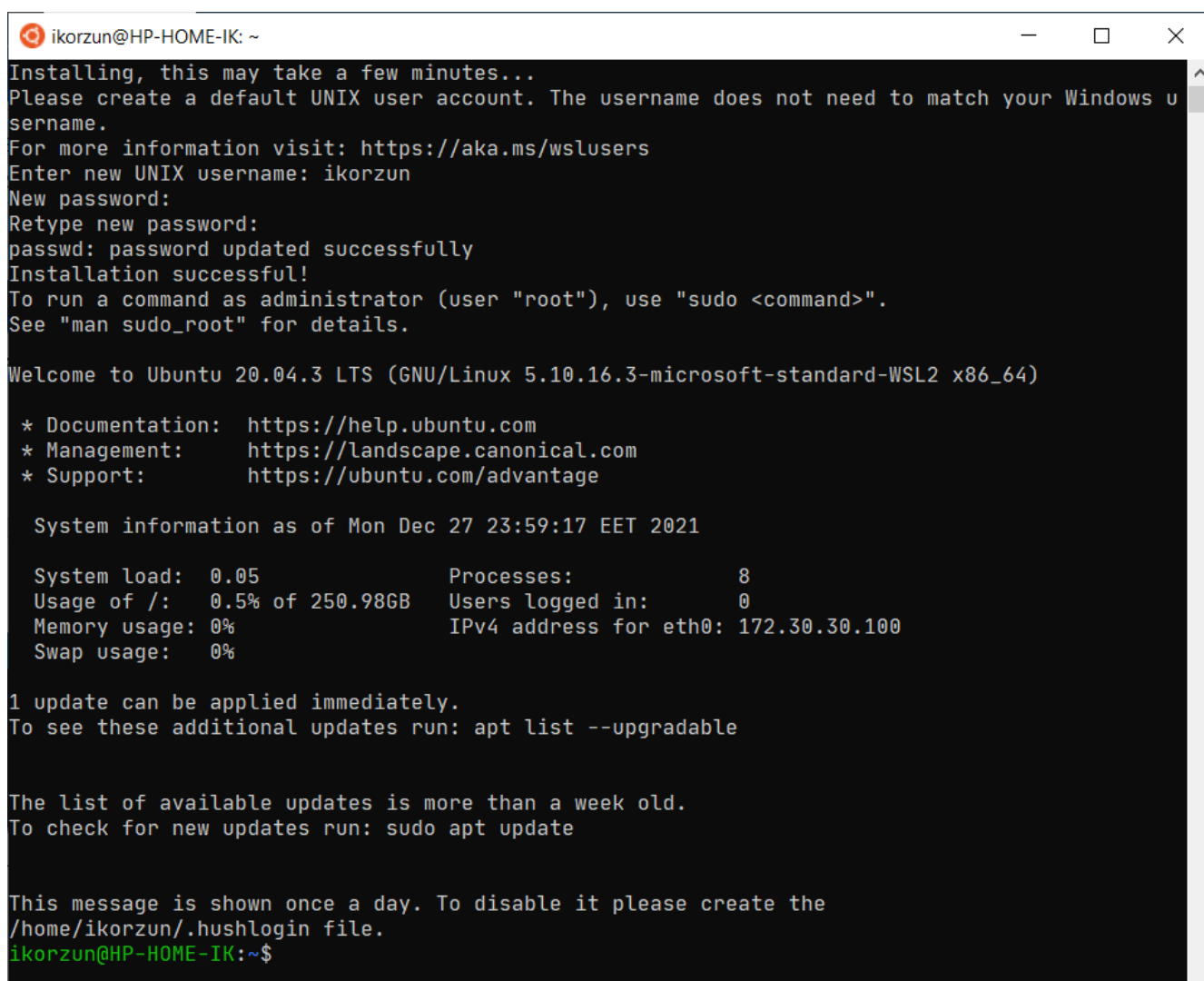
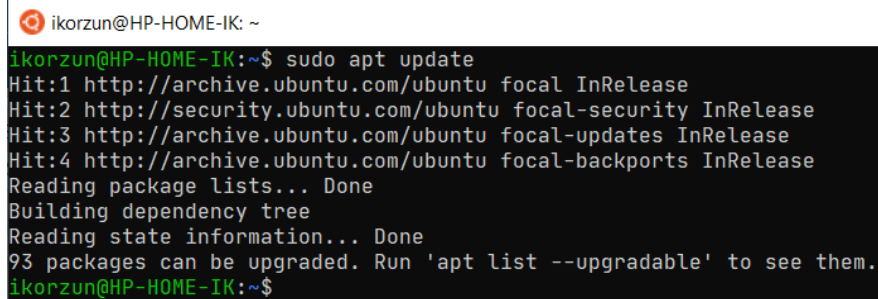


Рисунок 4. Створення профілю користувача та визначення паролю для нього

Встановлення Apache Hadoop у WSL

- оновлення індексу пакетів перед встановленням ПЗ в Linux

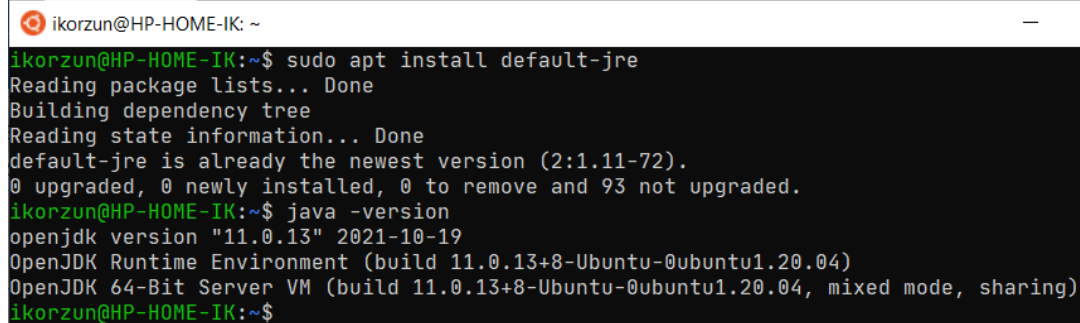
```
$ sudo apt update
```



```
ikorzun@HP-HOME-IK: ~  
ikorzun@HP-HOME-IK:~$ sudo apt update  
Hit:1 http://archive.ubuntu.com/ubuntu focal InRelease  
Hit:2 http://security.ubuntu.com/ubuntu focal-security InRelease  
Hit:3 http://archive.ubuntu.com/ubuntu focal-updates InRelease  
Hit:4 http://archive.ubuntu.com/ubuntu focal-backports InRelease  
Reading package lists... Done  
Building dependency tree  
Reading state information... Done  
93 packages can be upgraded. Run 'apt list --upgradable' to see them.  
ikorzun@HP-HOME-IK:~$
```

- встановлення актуальної версії Java

```
$ sudo apt install default-jre
```



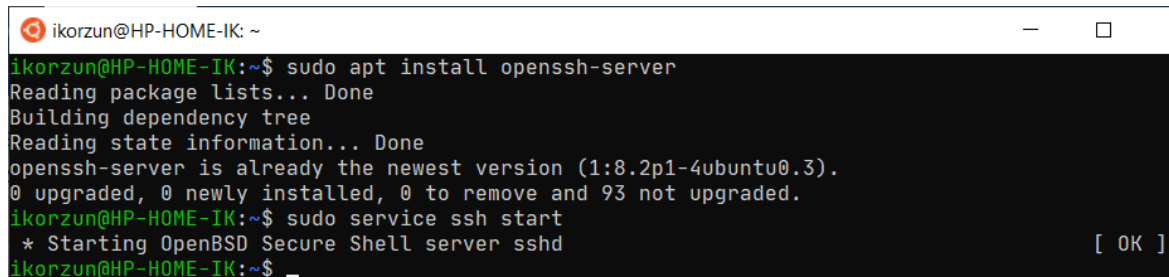
```
ikorzun@HP-HOME-IK: ~  
ikorzun@HP-HOME-IK:~$ sudo apt install default-jre  
Reading package lists... Done  
Building dependency tree  
Reading state information... Done  
default-jre is already the newest version (2:1.11-72).  
0 upgraded, 0 newly installed, 0 to remove and 93 not upgraded.  
ikorzun@HP-HOME-IK:~$ java -version  
openjdk version "11.0.13" 2021-10-19  
OpenJDK Runtime Environment (build 11.0.13+8-Ubuntu-0ubuntu1.20.04)  
OpenJDK 64-Bit Server VM (build 11.0.13+8-Ubuntu-0ubuntu1.20.04, mixed mode, sharing)  
ikorzun@HP-HOME-IK:~$
```

- встановлення OpenSSH

```
$ sudo apt remove openssh-server
```

```
$ sudo apt install openssh-server
```

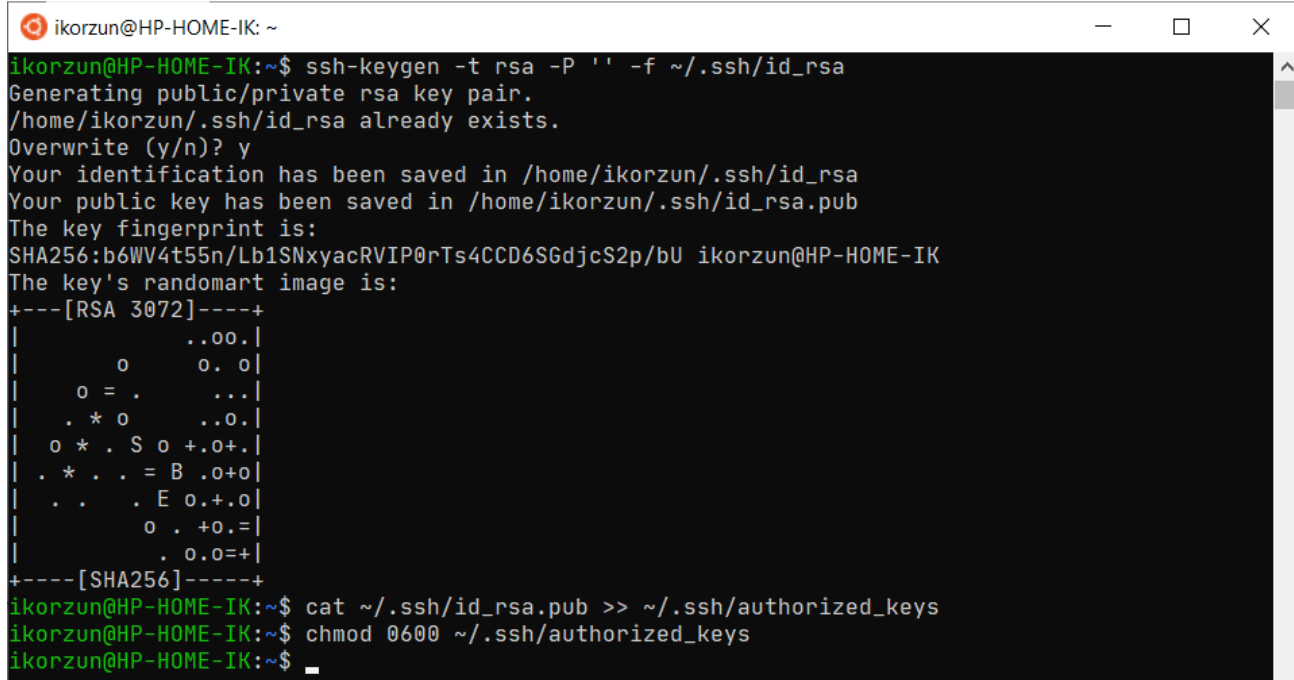
```
$ sudo service ssh start
```



```
ikorzun@HP-HOME-IK: ~  
ikorzun@HP-HOME-IK:~$ sudo apt install openssh-server  
Reading package lists... Done  
Building dependency tree  
Reading state information... Done  
openssh-server is already the newest version (1:8.2p1-4ubuntu0.3).  
0 upgraded, 0 newly installed, 0 to remove and 93 not upgraded.  
ikorzun@HP-HOME-IK:~$ sudo service ssh start  
* Starting OpenBSD Secure Shell server sshd  
ikorzun@HP-HOME-IK:~$
```

- створення ключів для локального доступу

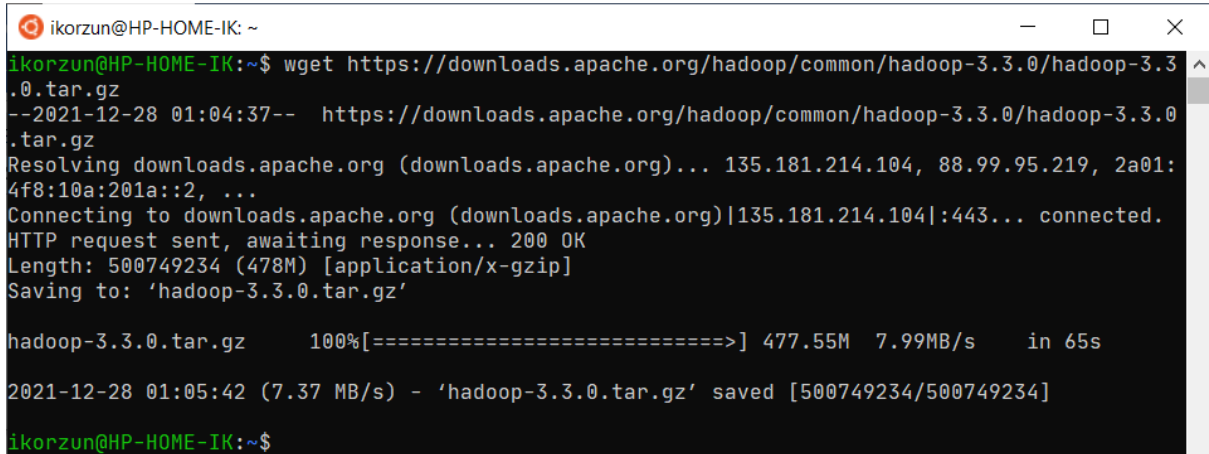
```
$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
$ chmod 0600 ~/.ssh/authorized_keys
```



```
ikorzun@HP-HOME-IK: ~
ikorzun@HP-HOME-IK:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
/home/ikorzun/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /home/ikorzun/.ssh/id_rsa
Your public key has been saved in /home/ikorzun/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:b6WV4t55n/Lb1SNxyacRVIP0rTs4CCD6SGdjcS2p/bU ikorzun@HP-HOME-IK
The key's randomart image is:
+---[RSA 3072]---+
|                .oo. |
|               o  o. |
|              o = .   |
|             . * o    |
|            o * . S o +.o+ |
|           . * . . = B .o+o |
|          . . . E o.+o. |
|         o . +o.= |
|        . o.o=+ |
+-----[SHA256]-----+
ikorzun@HP-HOME-IK:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
ikorzun@HP-HOME-IK:~$ chmod 0600 ~/.ssh/authorized_keys
ikorzun@HP-HOME-IK:~$
```

- завантаження актуальної збірки Apache Hadoop

```
$ wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
```



```
ikorzun@HP-HOME-IK: ~
ikorzun@HP-HOME-IK:~$ wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
--2021-12-28 01:04:37-- https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 135.181.214.104, 88.99.95.219, 2a01:4f8:10a:201a::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 500749234 (478M) [application/x-gzip]
Saving to: 'hadoop-3.3.0.tar.gz'

hadoop-3.3.0.tar.gz      100%[=====>] 477.55M  7.99MB/s   in 65s

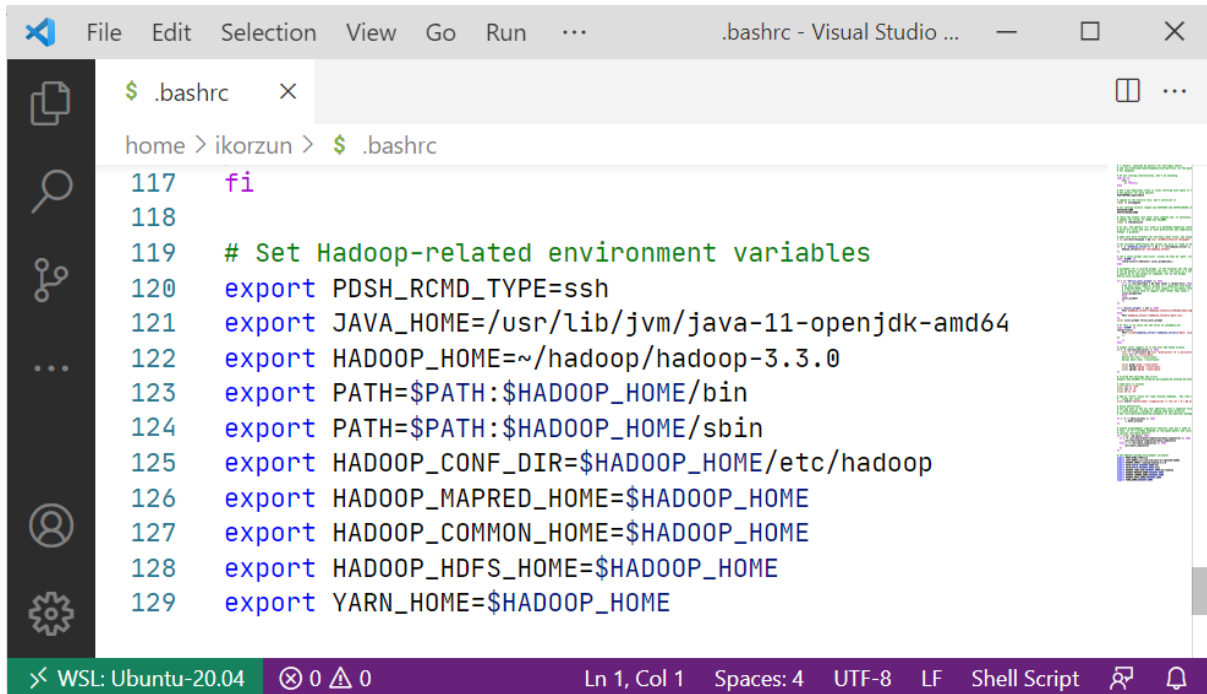
2021-12-28 01:05:42 (7.37 MB/s) - 'hadoop-3.3.0.tar.gz' saved [500749234/500749234]

ikorzun@HP-HOME-IK:~$
```

```
$ mkdir ~/hadoop
$ tar -xvzf hadoop-3.3.0.tar.gz -C ~/hadoop
```

— ВИЗНАЧЕННЯ СИСТЕМНИХ ЗМІННИХ

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

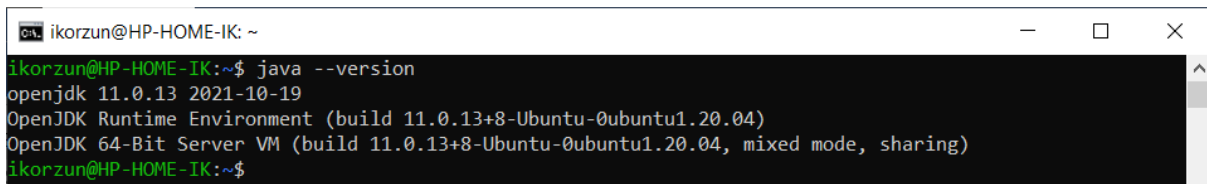


```
home > ikorzun > $ .bashrc
117 fi
118
119 # Set Hadoop-related environment variables
120 export PDSH_RCMD_TYPE=ssh
121 export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
122 export HADOOP_HOME=~/.hadoop/hadoop-3.3.0
123 export PATH=$PATH:$HADOOP_HOME/bin
124 export PATH=$PATH:$HADOOP_HOME/sbin
125 export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
126 export HADOOP_MAPRED_HOME=$HADOOP_HOME
127 export HADOOP_COMMON_HOME=$HADOOP_HOME
128 export HADOOP_HDFS_HOME=$HADOOP_HOME
129 export YARN_HOME=$HADOOP_HOME
```

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

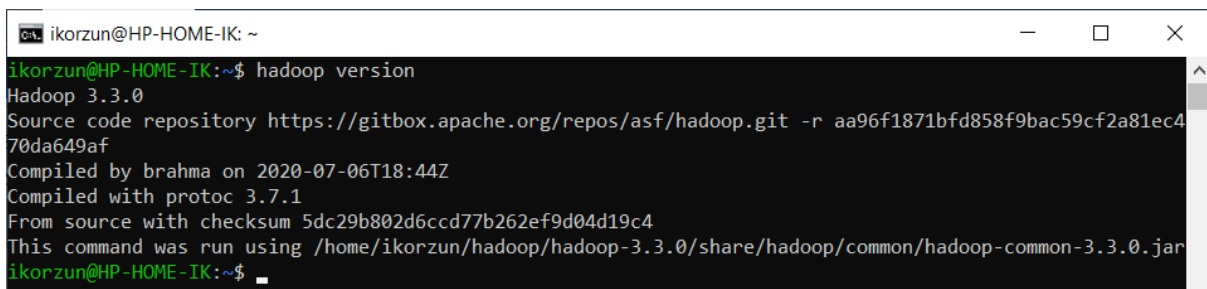
— якщо все зроблено правильно бачимо наступну відповідь

```
$ java -version
```



```
ikorzun@HP-HOME-IK: ~
ikorzun@HP-HOME-IK:~$ java --version
openjdk 11.0.13 2021-10-19
OpenJDK Runtime Environment (build 11.0.13+8-Ubuntu-0ubuntu1.20.04)
OpenJDK 64-Bit Server VM (build 11.0.13+8-Ubuntu-0ubuntu1.20.04, mixed mode, sharing)
ikorzun@HP-HOME-IK:~$
```

```
$ hadoop version
```

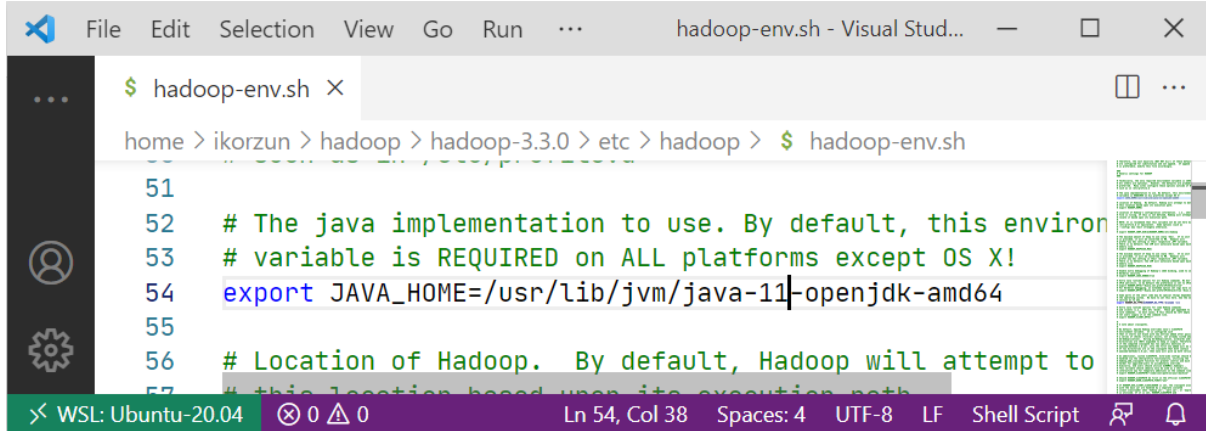


```
ikorzun@HP-HOME-IK: ~
ikorzun@HP-HOME-IK:~$ hadoop version
Hadoop 3.3.0
Source code repository https://gitbox.apache.org/repos/asf/hadoop.git -r aa96f1871bfd858f9bac59cf2a81ec470da649af
Compiled by brahma on 2020-07-06T18:44Z
Compiled with protoc 3.7.1
From source with checksum 5dc29b802d6ccd77b262ef9d04d19c4
This command was run using /home/ikorzun/hadoop/hadoop-3.3.0/share/hadoop/common/hadoop-common-3.3.0.jar
ikorzun@HP-HOME-IK:~$
```

Налаштування Apache Hadoop у WSL

– hadoop-env.sh

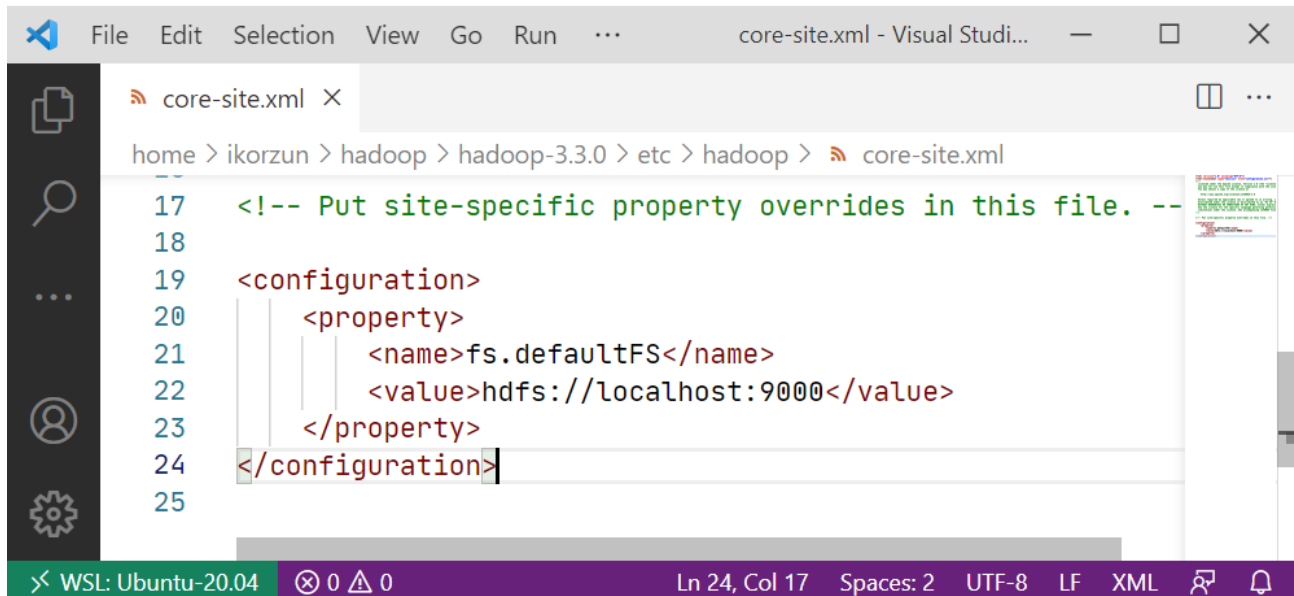
```
$ code $HADOOP_HOME/etc/hadoop/hadoop-env.sh
```



```
File Edit Selection View Go Run ... hadoop-env.sh - Visual Stud...
hadoop-env.sh x
home > ikorzun > hadoop > hadoop-3.3.0 > etc > hadoop > $ hadoop-env.sh
51
52 # The java implementation to use. By default, this environ
53 # variable is REQUIRED on ALL platforms except OS X!
54 export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
55
56 # Location of Hadoop. By default, Hadoop will attempt to
57 # this location based upon its execution path
Ln 54, Col 38 Spaces: 4 UTF-8 LF Shell Script
```

– core-site.xml

```
$ code $HADOOP_HOME/etc/hadoop/core-site.xml
```



```
File Edit Selection View Go Run ... core-site.xml - Visual Studi...
core-site.xml x
home > ikorzun > hadoop > hadoop-3.3.0 > etc > hadoop > core-site.xml
17 <!-- Put site-specific property overrides in this file. --
18
19 <configuration>
20   <property>
21     <name>fs.defaultFS</name>
22     <value>hdfs://localhost:9000</value>
23   </property>
24 </configuration>
25
Ln 24, Col 17 Spaces: 2 UTF-8 LF XML
```


– hdfs-site.xml

```
$ code $HADOOP_HOME/etc/hadoop/hdfs-site.xml
```

```
File Edit Selection View Go Run ... hdfs-site.xml - Visual Studi...  
hdfs-site.xml X  
home > ikorzun > hadoop > hadoop-3.3.0 > etc > hadoop > hdfs-site.xml  
18  
19 <configuration>  
20   <property>  
21     <name>dfs.replication</name>  
22     <value>1</value>  
23   </property>  
24 </configuration>  
25  
WSL: Ubuntu-20.04 0 0 Ln 24, Col 17 Spaces: 2 UTF-8 LF XML
```

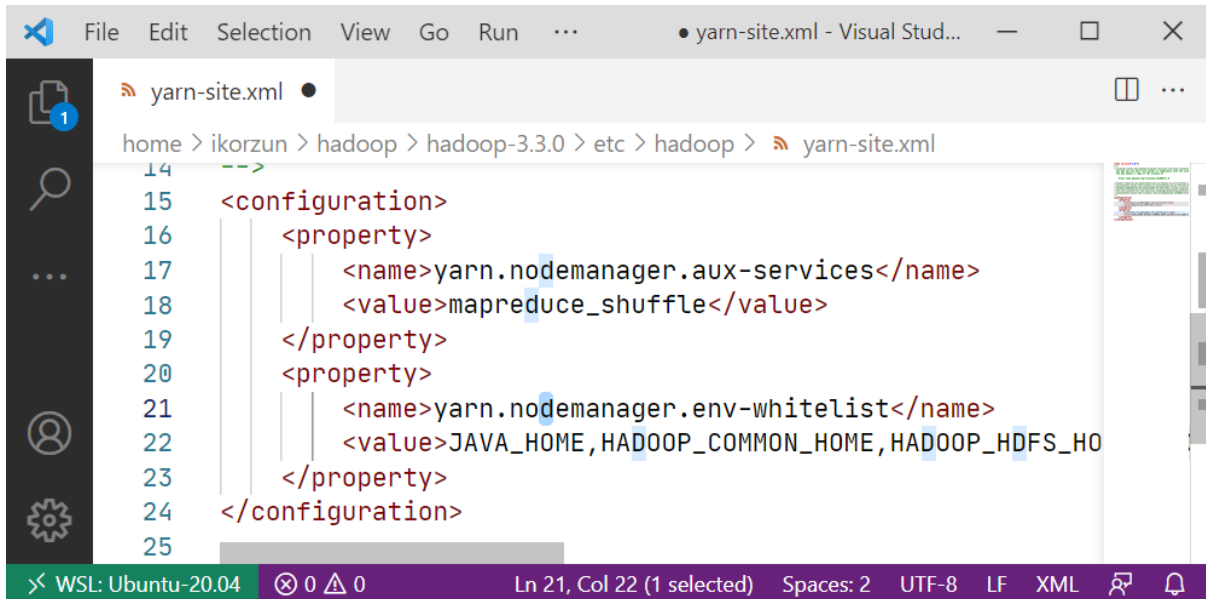
– mapred-site.xml

```
$ code $HADOOP_HOME/etc/hadoop/mapred-site.xml
```

```
File Edit Selection View Go Run ... mapred-site.xml - Visual Stu...  
mapred-site.xml X  
home > ikorzun > hadoop > hadoop-3.3.0 > etc > hadoop > mapred-site.xml  
19 <configuration>  
20   <property>  
21     <name>mapreduce.framework.name</name>  
22     <value>yarn</value>  
23   </property>  
24   <property>  
25     <name>mapreduce.application.classpath</name>  
26     <value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/  
27   </property>  
28 </configuration>  
29  
WSL: Ubuntu-20.04 0 0 Ln 28, Col 17 Spaces: 2 UTF-8 LF XML
```

– yarn-site.xml

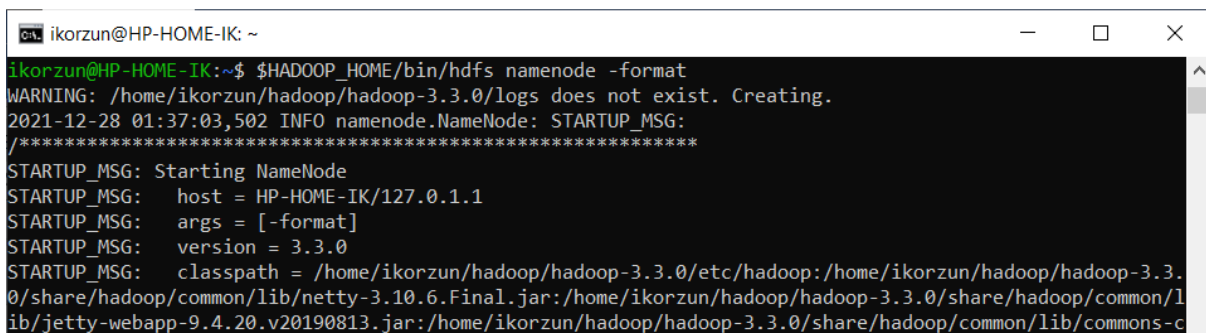
```
$ code $HADOOP_HOME/etc/hadoop/yarn-site.xml
```



```
home > ikorzun > hadoop > hadoop-3.3.0 > etc > hadoop > yarn-site.xml
14  -->
15  <configuration>
16    <property>
17      <name>yarn.nodemanager.aux-services</name>
18      <value>mapreduce_shuffle</value>
19    </property>
20    <property>
21      <name>yarn.nodemanager.env-whitelist</name>
22      <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HO
23    </property>
24  </configuration>
25
```

– ініціалізація файлової системи DFS

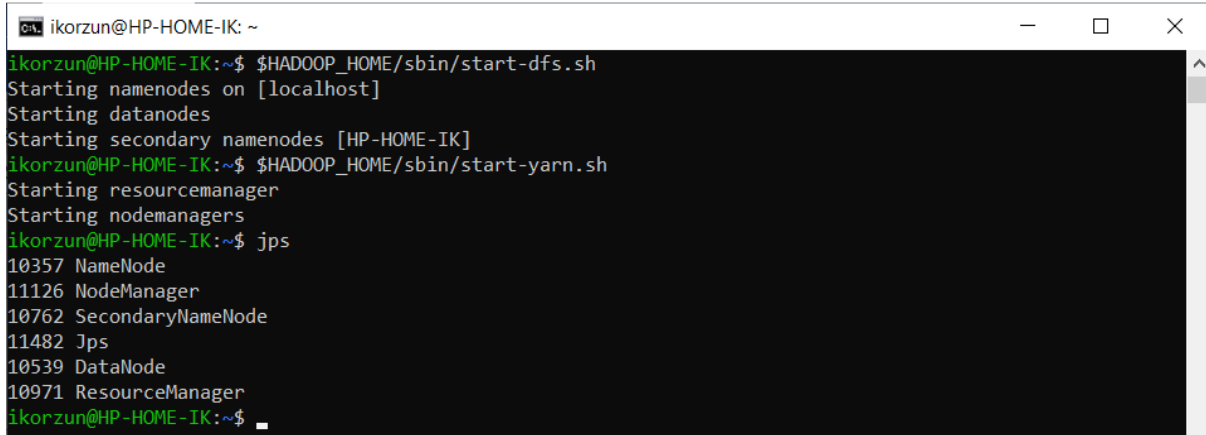
```
$ $HADOOP_HOME/bin/hdfs namenode -format
```



```
ikorzun@HP-HOME-IK: ~
ikorzun@HP-HOME-IK:~$ $HADOOP_HOME/bin/hdfs namenode -format
WARNING: /home/ikorzun/hadoop/hadoop-3.3.0/logs does not exist. Creating.
2021-12-28 01:37:03,502 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = HP-HOME-IK/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.3.0
STARTUP_MSG: classpath = /home/ikorzun/hadoop/hadoop-3.3.0/etc/hadoop:/home/ikorzun/hadoop/hadoop-3.3.0/share/hadoop/common/lib/netty-3.10.6.Final.jar:/home/ikorzun/hadoop/hadoop-3.3.0/share/hadoop/common/lib/jetty-webapp-9.4.20.v20190813.jar:/home/ikorzun/hadoop/hadoop-3.3.0/share/hadoop/common/lib/commons-c
```

– запуск сервісів Hadoop

```
$HADOOP_HOME/sbin/start-dfs.sh  
$HADOOP_HOME/sbin/start-yarn.sh
```

A terminal window titled 'ikorzun@HP-HOME-IK: ~' with standard window controls. It displays the output of the Hadoop startup scripts. The first command, '\$HADOOP_HOME/sbin/start-dfs.sh', shows the process of starting namenodes on localhost, datanodes, and secondary namenodes. The second command, '\$HADOOP_HOME/sbin/start-yarn.sh', shows the process of starting the ResourceManager, NodeManagers, and Jps. The final command, 'jps', lists the running processes with their PIDs: NameNode (10357), NodeManager (11126), SecondaryNameNode (10762), Jps (11482), DataNode (10539), and ResourceManager (10971).

```
ikorzun@HP-HOME-IK:~$ $HADOOP_HOME/sbin/start-dfs.sh  
Starting namenodes on [localhost]  
Starting datanodes  
Starting secondary namenodes [HP-HOME-IK]  
ikorzun@HP-HOME-IK:~$ $HADOOP_HOME/sbin/start-yarn.sh  
Starting resourcemanager  
Starting nodemanagers  
ikorzun@HP-HOME-IK:~$ jps  
10357 NameNode  
11126 NodeManager  
10762 SecondaryNameNode  
11482 Jps  
10539 DataNode  
10971 ResourceManager  
ikorzun@HP-HOME-IK:~$
```

Налаштування Python

```
$ sudo apt-get install python3.10  
$ sudo apt-get install python3-pip  
$ python3 -m pip install mrjob
```

Реалізація обраної задачі за допомогою бібліотеки MrJob мовою Python

```
from mrjob.job import MRJob
from mrjob.step import MRStep
import re

WORD_RE = re.compile(r"[\w']+")

class MostUsedWord(MRJob):

    def steps(self):
        return [
            MRStep(mapper=self.mapper_get_words,
                  combiner=self.combiner_count_words,
                  reducer=self.reducer_count_words),
            MRStep(reducer=self.reducer_find_max_word)
        ]

    def mapper_get_words(self, _, line):
        # yield each word in the line
        for word in WORD_RE.findall(line):
            yield (word.lower(), 1)

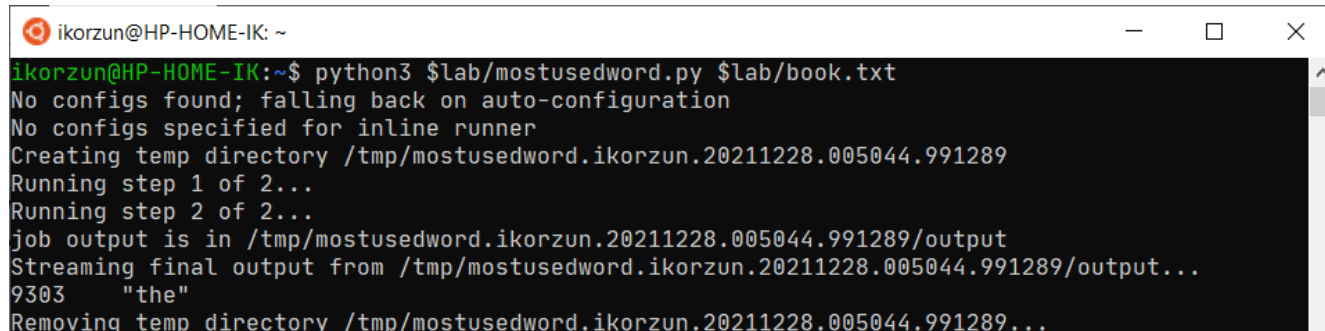
    def combiner_count_words(self, word, counts):
        # optimization: sum the words we've seen so far
        yield (word, sum(counts))

    def reducer_count_words(self, word, counts):
        # send all (num_occurrences, word) pairs to the same reducer.
        # num_occurrences is so we can easily use Python's max() function.
        yield None, (sum(counts), word)

    # discard the key; it is just None
    def reducer_find_max_word(self, _, word_count_pairs):
        # each item of word_count_pairs is (count, word),
        # so yielding one results in key=counts, value=word
        yield max(word_count_pairs)

if __name__ == '__main__':
    MostUsedWord.run()
```

— тестування програми (без платформи Hadoop)



A terminal window titled 'ikorzun@HP-HOME-IK: ~' showing the execution of the program. The command executed is 'python3 \$lab/mostusedword.py \$lab/book.txt'. The output shows the program running two steps and producing the final output '9303 "the"'. The terminal text is as follows:

```
ikorzun@HP-HOME-IK:~$ python3 $lab/mostusedword.py $lab/book.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mostusedword.ikorzun.20211228.005044.991289
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mostusedword.ikorzun.20211228.005044.991289/output
Streaming final output from /tmp/mostusedword.ikorzun.20211228.005044.991289/output...
9303      "the"
Removing temp directory /tmp/mostusedword.ikorzun.20211228.005044.991289...
```

– тестування програми (із залученням Hadoop)

```
ikorzun@HP-HOME-IK: ~  
ikorzun@HP-HOME-IK:~$ python3 $lab/mostusedword.py -r hadoop $lab/book.txt  
No configs found; falling back on auto-configuration  
No configs specified for hadoop runner  
Looking for hadoop binary in /home/ikorzun/hadoop/hadoop-3.3.0/bin...  
Found hadoop binary: /home/ikorzun/hadoop/hadoop-3.3.0/bin/hadoop  
Using Hadoop version 3.3.0  
Looking for Hadoop streaming jar in /home/ikorzun/hadoop/hadoop-3.3.0...  
Found Hadoop streaming jar: /home/ikorzun/hadoop/hadoop-3.3.0/share/hadoop/tools/lib/hadoop-streaming-3.3.0.jar  
Creating temp directory /tmp/mostusedword.ikorzun.20211228.005105.391877  
uploading working dir files to hdfs:///user/ikorzun/tmp/mrjob/mostusedword.ikorzun.20211228.005105.391877/files/wd...  
Copying other local files to hdfs:///user/ikorzun/tmp/mrjob/mostusedword.ikorzun.20211228.005105.391877/files/  
Running step 1 of 2...  
packageJobJar: [/tmp/hadoop-unjar15556356832670095766/] [] /tmp/streamjob2933108657374787866.jar tmpDir=null  
Connecting to ResourceManager at /0.0.0.0:8032  
Connecting to ResourceManager at /0.0.0.0:8032  
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ikorzun/.staging/job_1640649225823_0001  
Total input files to process : 1  
number of splits:2  
Submitting tokens for job: job_1640649225823_0001  
Executing with tokens: []  
resource-types.xml not found  
Unable to find 'resource-types.xml'.  
Submitted application application_1640649225823_0001  
The url to track the job: http://HP-HOME-IK.localdomain:8088/proxy/application_1640649225823_0001/  
Running job: job_1640649225823_0001  
Job job_1640649225823_0001 running in uber mode : false  
map 0% reduce 0%  
map 100% reduce 0%  
map 100% reduce 100%  
Job job_1640649225823_0001 completed successfully  
Output directory: hdfs:///user/ikorzun/tmp/mrjob/mostusedword.ikorzun.20211228.005105.391877/step-output/0000  
Counters: 54  
File Input Format Counters  
Bytes Read=678666  
File Output Format Counters  
Bytes Written=206926  
  
Shuffle Errors  
BAD_ID=0  
CONNECTION=0  
IO_ERROR=0  
WRONG_LENGTH=0  
WRONG_MAP=0  
WRONG_REDUCE=0  
job output is in hdfs:///user/ikorzun/tmp/mrjob/mostusedword.ikorzun.20211228.005105.391877/output  
Streaming final output from hdfs:///user/ikorzun/tmp/mrjob/mostusedword.ikorzun.20211228.005105.391877/output...  
9303 "the"  
Removing HDFS temp directory hdfs:///user/ikorzun/tmp/mrjob/mostusedword.ikorzun.20211228.005105.391877...  
Removing temp directory /tmp/mostusedword.ikorzun.20211228.005105.391877...  
ikorzun@HP-HOME-IK:~$
```

Висновок

У результаті виконання лабораторної роботи реалізовано рішення для підрахунку слів з використанням підходу MapReduce на платформі Apache Hadoop. У ході роботи вивчено особливості встановлення, налаштування та застосування платформи та окремо методи створення рішень із залучення Hadoop до обчислень на великих даних за допомогою мови Python, хоча для даного інструменту Java є більш нативним методом розробки.