

Họ và tên: Bùi Thiện Nhân

MSSV: 2274802010592

LAB 07 NHẬP MÔN PHÂN TÍCH DỮ LIỆU LỚN

1. Đăng nhập server qua SSH:

```
hadoop@hadoop-2274802010592:~/iDragonCloud$
```

2. Kiểm tra phiên bản Hadoop:

```
hadoop@hadoop-2274802010592:~/iDragonCloud$ hadoop version
Hadoop 3.4.0
```

3. Kiểm tra JDK:

```
hadoop@hadoop-2274802010592:~/iDragonCloud$ java -version
openjdk version "11.0.26" 2025-01-21
OpenJDK Runtime Environment (build 11.0.26+4-post-Ubuntu-1ubuntu122.04)
OpenJDK 64-Bit Server VM (build 11.0.26+4-post-Ubuntu-1ubuntu122.04, mixed mode, sharing)
hadoop@hadoop-2274802010592:~/iDragonCloud$
```

4.2 Kiểm tra và cài đặt thư viện

1. Kiểm tra thư viện Hadoop:

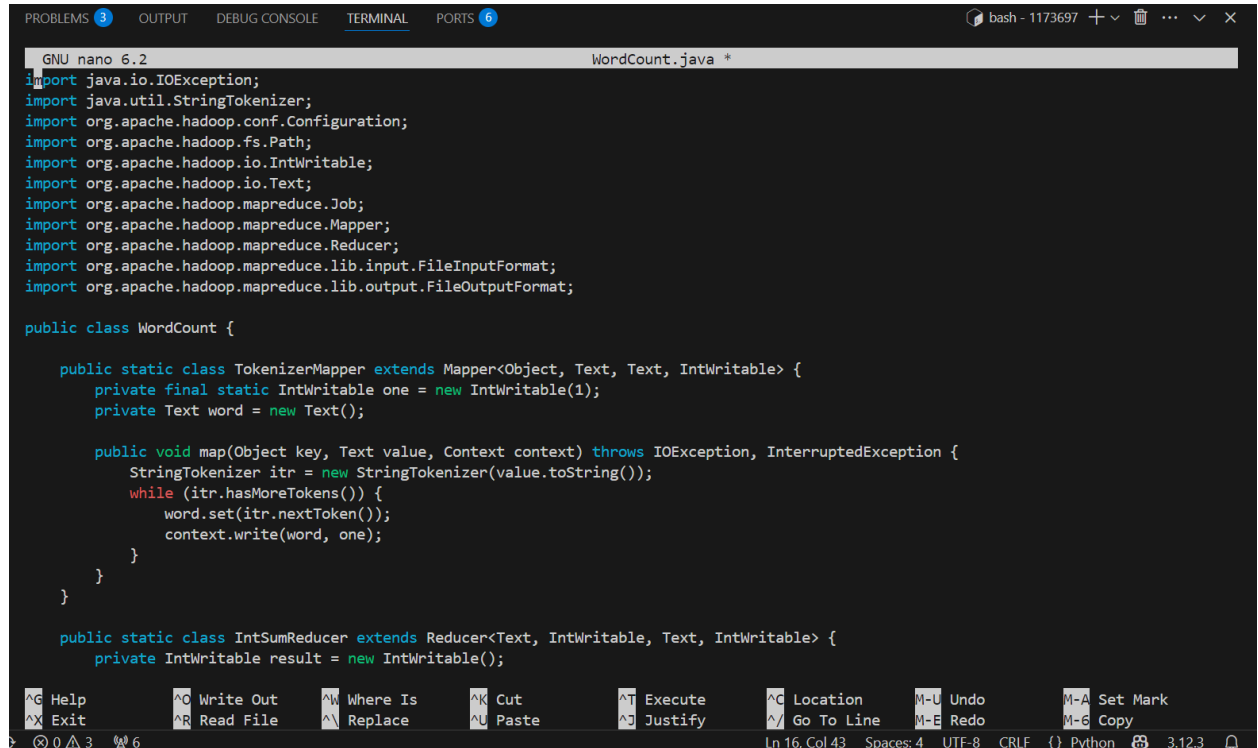
```
hadoop@hadoop-2274802010592:~/iDragonCloud$ ls / usr / local / hadoop / share / hadoop / common /
ls: cannot access 'usr': No such file or directory
ls: cannot access 'local': No such file or directory
ls: cannot access 'hadoop': No such file or directory
ls: cannot access 'share': No such file or directory
ls: cannot access 'hadoop': No such file or directory
ls: cannot access 'common': No such file or directory
/:
bin boot dev etc home lib lib32 lib64 libx32 media mnt opt proc root run sbin srv sys tmp usr var
/:
bin boot dev etc home lib lib32 lib64 libx32 media mnt opt proc root run sbin srv sys tmp usr var
/:
bin boot dev etc home lib lib32 lib64 libx32 media mnt opt proc root run sbin srv sys tmp usr var
/:
bin boot dev etc home lib lib32 lib64 libx32 media mnt opt proc root run sbin srv sys tmp usr var
/:
bin boot dev etc home lib lib32 lib64 libx32 media mnt opt proc root run sbin srv sys tmp usr var
/:
bin boot dev etc home lib lib32 lib64 libx32 media mnt opt proc root run sbin srv sys tmp usr var
hadoop@hadoop-2274802010592:~/iDragonCloud$
```

4.3 Viết chương trình Word Count

1. Tạo thư mục làm việc:

```
hadoop@hadoop-2274802010592:~/iDragonCloud$ mkdir ~/wordcount_lab
hadoop@hadoop-2274802010592:~/iDragonCloud$ cd ~/wordcount_lab
hadoop@hadoop-2274802010592:~/wordcount_lab$
```

2. Tạo file WordCount.java: (Mã nguồn giữ nguyên, không chỉnh sửa vì đã đủ rõ ràng)



```
GNU nano 6.2 WordCount.java *
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {

    public static class TokenizerMapper extends Mapper<Object, Text, Text, IntWritable> {
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
            StringTokenizer itr = new StringTokenizer(value.toString());
            while (itr.hasMoreTokens()) {
                word.set(itr.nextToken());
                context.write(word, one);
            }
        }
    }

    public static class IntSumReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
        private IntWritable result = new IntWritable();
    }
}
```

4. Biên dịch (ngắt dòng lệnh dài):

```

hadoop@hadoop-2274802010592:~/wordcount_lab$ javac -cp $HADOOP_HOME/share/hadoop/common/hadoop-common-*.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-core-*.jar WordCount.java
WordCount.java:3: error: package org.apache.hadoop.conf does not exist
import org.apache.hadoop.conf.Configuration;
                        ^
WordCount.java:4: error: package org.apache.hadoop.fs does not exist
import org.apache.hadoop.fs.Path;
                        ^
WordCount.java:5: error: package org.apache.hadoop.io does not exist
import org.apache.hadoop.io.IntWritable;
                        ^
WordCount.java:6: error: package org.apache.hadoop.io does not exist
import org.apache.hadoop.io.Text;
                        ^
WordCount.java:7: error: package org.apache.hadoop.mapreduce does not exist
import org.apache.hadoop.mapreduce.Job;
                        ^
WordCount.java:8: error: package org.apache.hadoop.mapreduce does not exist
import org.apache.hadoop.mapreduce.Mapper;
                        ^
WordCount.java:9: error: package org.apache.hadoop.mapreduce does not exist
import org.apache.hadoop.mapreduce.Reducer;
                        ^
WordCount.java:10: error: package org.apache.hadoop.mapreduce.lib.input does not exist
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
                        ^
WordCount.java:11: error: package org.apache.hadoop.mapreduce.lib.output does not exist
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
                        ^
WordCount.java:15: error: cannot find symbol
    public static class TokenizerMapper extends Mapper<Object, Text, Text, IntWritable> {
                                           ^

```

5. Tạo file JAR:

```

hadoop@hadoop-2274802010592:~/wordcount_lab$ ls -la WordCount.java
-rw-r--r-- 1 hadoop fitlab 2114 Apr  1 10:48 WordCount.java
hadoop@hadoop-2274802010592:~/wordcount_lab$ javac -cp $(hadoop classpath) WordCount.java
hadoop@hadoop-2274802010592:~/wordcount_lab$ ls -la WordCount*.class
-rw-r--r-- 1 hadoop fitlab 1739 Apr  1 10:51 'WordCount$IntSumReducer.class'
-rw-r--r-- 1 hadoop fitlab 1736 Apr  1 10:51 'WordCount$TokenizerMapper.class'
-rw-r--r-- 1 hadoop fitlab 1452 Apr  1 10:51 WordCount.class
hadoop@hadoop-2274802010592:~/wordcount_lab$ jar cf wordcount.jar WordCount*.class
hadoop@hadoop-2274802010592:~/wordcount_lab$ ls -la wordcount.jar
-rw-r--r-- 1 hadoop fitlab 3050 Apr  1 10:52 wordcount.jar
hadoop@hadoop-2274802010592:~/wordcount_lab$

```

4.4 Chuẩn bị dữ liệu đầu vào

1. Tạo file input.txt:

```

hadoop@hadoop-2274802010592:~/wordcount_lab$ echo "Hadoop is great. Hadoop is easy." > input.txt
hadoop@hadoop-2274802010592:~/wordcount_lab$

```

2. Tải lên HDFS:

```

hadoop@hadoop-2274802010592:~/wordcount_lab$ hdfs dfs -mkdir -p /user/hadoop/input
mkdir: Call From hadoop-2274802010592/172.23.0.105 to localhost:9000 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
hadoop@hadoop-2274802010592:~/wordcount_lab$ hdfs dfs -put input.txt /user/hadoop/input
put: Call From hadoop-2274802010592/172.23.0.105 to localhost:9000 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
hadoop@hadoop-2274802010592:~/wordcount_lab$

```

4.5 Chạy chương trình

1. Thực thi job MapReduce (ngắt dòng lệnh dài):

```
hadoop@hadoop-2274802010592:~/wordcount_lab3$ hadoop jar wordcount.jar WordCount \
/user/hadoop/input /user/hadoop/output
2025-04-01 17:56:09,243 INFO Impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-04-01 17:56:09,329 INFO Impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-04-01 17:56:09,330 INFO Impl.MetricsSystemImpl: JobTracker metrics system started
Exception in thread "main" java.net.ConnectException: Call From hadoop-2274802010592/172.23.0.105 to localhost:9000 failed on connection ex
ception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
    at java.base/jdk.internal.reflect.NativeConstructorAccessorImpl.newInstance(Native Method)
    at java.base/jdk.internal.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
    at java.base/jdk.internal.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
    at java.base/java.lang.reflect.Constructor.newInstance(Constructor.java:490)
    at org.apache.hadoop.net.NetUtils.wrapWithMessage(NetUtils.java:948)
    at org.apache.hadoop.net.NetUtils.wrapException(NetUtils.java:863)
    at org.apache.hadoop.ipc.Client.getRpcResponse(Client.java:1588)
    at org.apache.hadoop.ipc.Client.call(Client.java:1529)
    at org.apache.hadoop.ipc.Client.call(Client.java:1426)
    at org.apache.hadoop.ipc.ProtobufRpcEngine2$Invoker.invoke(ProtobufRpcEngine2.java:258)
    at org.apache.hadoop.ipc.ProtobufRpcEngine2$Invoker.invoke(ProtobufRpcEngine2.java:139)
```

2. Kiểm tra kết quả:

```

at org.apache.hadoop.mapreduce.Job.submit(Job.java:1674)
at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1695)
at WordCount.main(WordCount.java:51)
at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at java.base/jdk.internal.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.base/java.lang.reflect.Method.invoke(Method.java:566)
at org.apache.hadoop.util.RunJar.run(RunJar.java:330)
at org.apache.hadoop.util.RunJar.main(RunJar.java:245)
Caused by: java.net.ConnectException: Connection refused
at java.base/sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
at java.base/sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:777)
at org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.java:205)
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:601)
at org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:668)
at org.apache.hadoop.ipc.Client$Connection.setupIOStreams(Client.java:789)
at org.apache.hadoop.ipc.Client$Connection.access$3800(Client.java:364)
at org.apache.hadoop.ipc.Client.getConnection(Client.java:1649)
at org.apache.hadoop.ipc.Client.call(Client.java:1473)
... 40 more
hadoop@hadoop-2274802010592:~/wordcount_lab$ hdfs dfs -cat /user/hadoop/output/part-r-00000
cat: Call From hadoop-2274802010592/172.23.0.195 to localhost:9000 failed on connection exception: java.net.ConnectException: Connection re
fused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
hadoop@hadoop-2274802010592:~/wordcount_lab$

```

4.6 Xóa thư mục (Tùy chọn)

```
hadoop@hadoop-2274802010592:~/wordcount_lab$ hdfs dfs - rm -r / user / hadoop / output
rm: Unknown command
Did you mean -rm? This command begins with a dash.
Usage: hadoop fs [generic options]
    [-appendToFile [-n] <localsrc> ... <dst>]
    [-cat [-ignoreCrc] <src> ...]
    [-checksum [-v] <src> ...]
    [-chgrp [-R] GROUP PATH...]
    [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
    [-chown [-R] [OWNER][[:GROUP]] PATH...]
    [-concat <target path> <src path> <src path> ...]
    [-copyFromLocal [-f] [-p] [-l] [-d] [-t <thread count>] [-q <thread pool queue size>] <localsrc> ... <dst>]
    [-copyToLocal [-f] [-p] [-crc] [-ignoreCrc] [-t <thread count>] [-q <thread pool queue size>] <src> ... <localdst>]
    [-count [-q] [-h] [-v] [-t <storage type>]] [-u] [-x] [-e] [-s] <path> ...]
    [-cp [-f] [-p] [-p[topax]] [-d] [-t <thread count>] [-q <thread pool queue size>] <src> ... <dst>]
    [-createSnapshot <snapshotDir> [<snapshotName>]]
    [-deleteSnapshot <snapshotDir> <snapshotName>]
    [-df [-h] [<path> ...]]
    [-du [-s] [-h] [-v] [-x] <path> ...]
    [-expunge [-immediate] [-fs <path>]]
    [-find <path> ... <expression> ...]
    [-fs <path> ... <expression> ...] [-t <thread count>] [-q <thread pool queue size>] [-x] [-e] [-s] <path> ...]
```