# 1. Set up

Vì đã thay đổi network để setup môi trường cho hệ thống hadoop multi node nên không kết nối với đường mạng hiện tại để tải các packages về. Vì vậy, ta nên tải từ window và mount vào hệ thống hadoop thông qua tools của VMWare

## Cài đặt Scala

Tải Scala về

### Trường hợp 1: máy có kết nối mạng bình thường

|  |
| --- |
| wget <https:////www.scala-lang.org/files/archive/scala-2.13.15.tgz> |



### 1.1.2 Trường hợp 2: máy không có kết nối mạng

Truy cập đường dẫn sau để tải gói packages phù hợp với linux

[Index of /files/archive (scala-lang.org)](https://www.scala-lang.org/files/archive/) 



Mount vào hệ thống linux

A screenshot of a computer

Description automatically generated

Vào Setting của máy ảo, bật chết độ cho phép share file và add đường dẫn vào

|  |
| --- |
| cd /mnt/hgfs/Scala |
| ls |
| cp scala-2.13.15.tgz /root/ |
| chmod 777 /root/scala-2.13.15.tgz |

Sau khi tải về đc package về 1 trong 2 cách trên

|  |
| --- |
| tar xvfz scala-2.13.15.tgz |
| mv scala-2.13.15 scala |
| mv scala /usr/local/share/scala |

A screen shot of a computer

Description automatically generated





Sau khi giả nén folder ra và di chuyển folder đến vị trí /usr/local/share/scala. Ở đường dẫn này thì các user sẽ chia sẽ file cho nhau, nếu user nào muốn dùng thì vào user đó khai báo biến môi trường của user đó

|  |
| --- |
| nano ~/.bashrc |

Thêm nội dung sau vào file

|  |
| --- |
| export SCALA\_HOME=/usr/local/share/scala  export PATH=$PATH:$SCALA\_HOME/bin |
| source ~/.bashrc |
| exec $SHELL |

A black background with red and green text

Description automatically generated

Sử dụng thì dùng lệnh

|  |
| --- |
| scala |

## Cài đặt Spark

### 1.2.1 Trường hợp 1: máy có kết nối mạng bình thường

|  |
| --- |
| wget https://dlcdn.apache.org/spark/spark-3.5.3/spark-3.5.3-bin-hadoop3.tgz |

Đem folder cài đặt spark vào thư mục tùy mình để dễ quản lý

### 1.2.2 Trường hợp 2: máy không có kết nối mạng

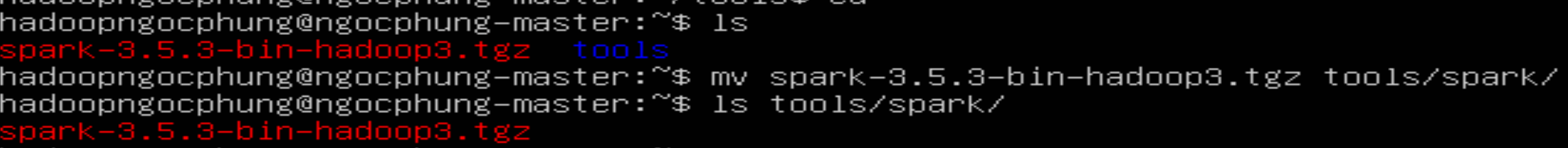
[Index of /spark/spark-3.5.3 (apache.org)](https://dlcdn.apache.org/spark/spark-3.5.3/)

Tải spark từ link này để có thể có pakage và mount vào hệ thống

|  |
| --- |
| cd /mnt/hgfs/Spark |
| ls |
| cp spark-3.5.3-bin-hadoop3.tgz /home/hadoopngocphung/tools/spark |
| chmod 777 home/hadoopngocphung/tools/spark/spark-3.5.3-bin-hadoop3.tgz |

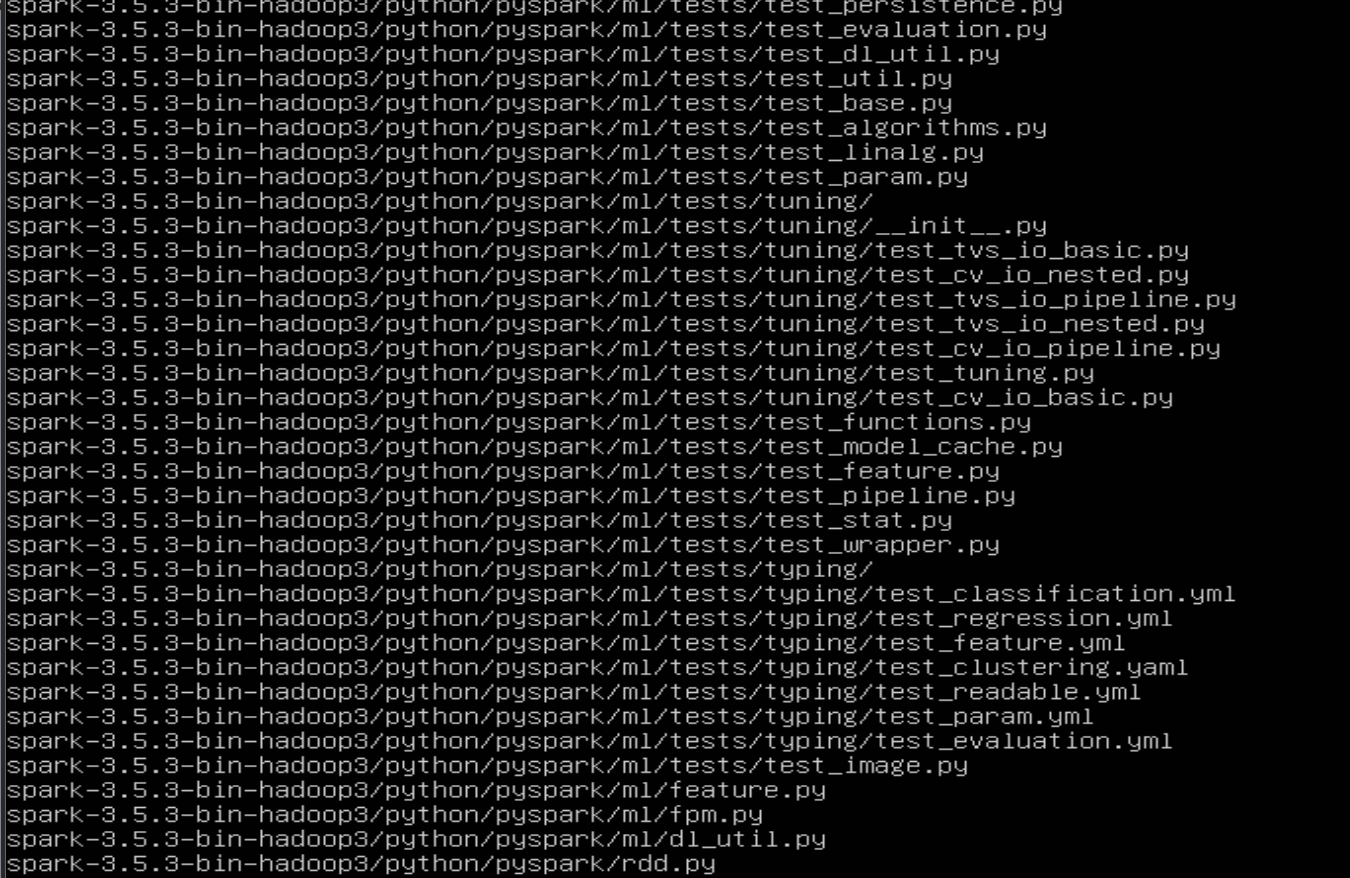
### 1.2.3 Cấu hình Spark

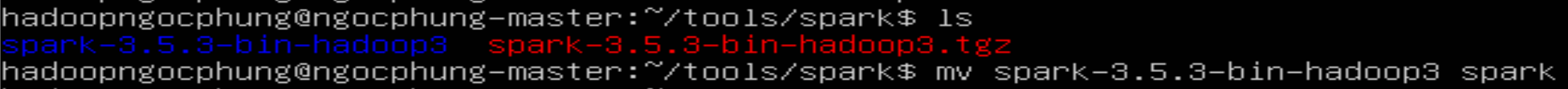
Sau khi có đc pakage ở 1 trong 2 cách trên



Giải nén tệp và đổi tên để dễ quản lý

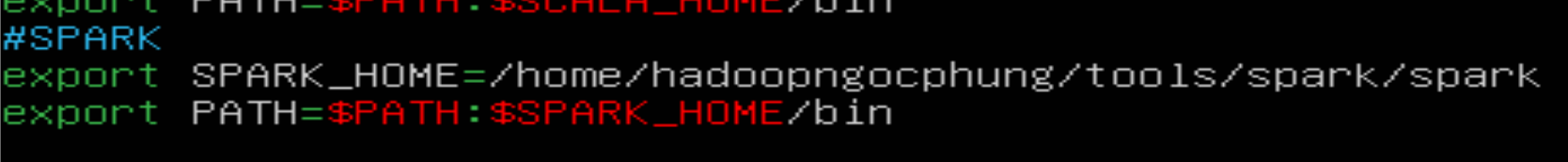
|  |
| --- |
| tar -xvzf spark-3.5.3-bin-hadoop3.tgz |
| mv spark-3.5.3-bin-hadoop3 spark |





Vào file bashrc để khai báo môi trường cho Spark

|  |
| --- |
| nano ~/.bashrc |
| export SPARK\_HOME=/home/hadoopngocphung/tools/spark/spark  export PATH=$PATH:$SPARK\_HOME/bin |



Khởi động spark

|  |
| --- |
| spark-shell |

A computer screen with white text

Description automatically generated

Thoát khỏi spark

|  |
| --- |
| :quit |

## Các đặt các jars

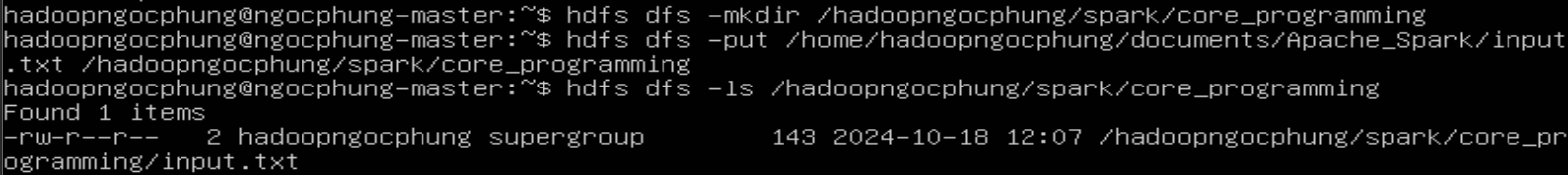
## 1.4 Chuẩn bị file

### 1.4.1 file input.txt

|  |
| --- |
| people are not as beautiful as they look,  as they walk or as they talk.  they are only as beautiful as they love,  as they care as they share. |

# 2. Core Programming

|  |
| --- |
| hdfs dfs -mkdir /hadoopngocphung/spark |
| hdfs dfs -mkdir /hadoopngocphung/spark/core\_programming |
| hdfs dfs -put /home/hadoopngocphung/documents/Apache\_Spark/input.txt /hadoopngocphung/spark/core\_programming |
| hdfs dfs -ls /hadoopngocphung/spark/core\_programming |



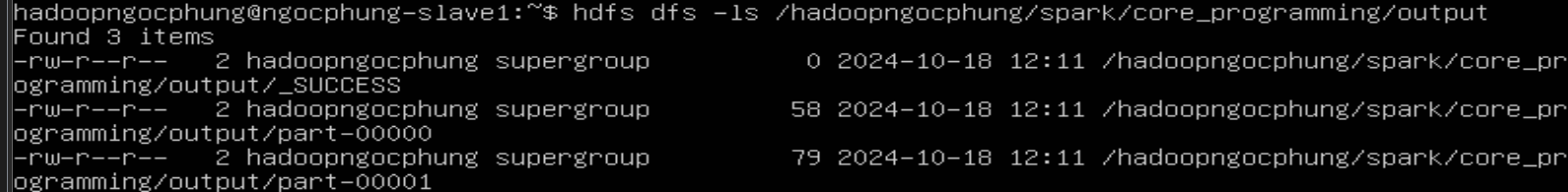
|  |
| --- |
| val inputfile = sc.textFile("hdfs://ngocphung-master:9000/hadoopngocphung/spark/core\_programming/input.txt") |
| val counts = inputfile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(\_+\_); |
| counts.toDebugString |
| counts.cache() |
| counts.saveAsTextFile("hdfs://ngocphung-master:9000/hadoopngocphung/spark/core\_programming/output") |

A computer screen shot of a computer code

Description automatically generated

Kiểm tra xem có folder output\_core hay không ?

|  |
| --- |
| hdfs dfs -mkdir /hadoopngocphung/spark/core\_programming/output |



Đọc file

A black screen with white text

Description automatically generated

A black screen with white text

Description automatically generated

# 3. Advanced Spark Programming

|  |
| --- |
| val broadcastVar = sc.broadcast(Array(1, 2, 3)) |
| val accum = sc.accumulator(0) |
| val accum = sc.longAccumulator("My Accumulator") |
| sc.parallelize(Array(1, 2, 3, 4)).foreach(x => accum += x) |
| accum.value |

Lỗi này xảy ra vì phương thức accumulator đã bị **loại bỏ** (deprecated) trong các phiên bản mới của Apache Spark. Trong các phiên bản mới hơn (Spark 2.x trở lên), accumulator đã được thay thế bằng LongAccumulator hoặc DoubleAccumulator từ gói org.apache.spark.util.

A screenshot of a computer program

Description automatically generated