

Spark : Installation et Configuration

BIG-DATA

El-Houssein Hamoudi C20536



Plan de la Présentation

1. Introduction
2. Étapes d'Installation et Configuration
3. Implémentation PySpark
4. Implémentation Scala
5. Démonstration
6. Conclusion

Introduction

- Cette présentation documente l'installation et la configuration d'Apache Spark (version 3.5.0) dans un environnement Ubuntu Server virtualisé. L'objectif principal était de déployer un cluster Spark mono-nœud et de valider son bon fonctionnement via des exemples concrets en PySpark (Python) et Scala. Les étapes clés comprennent :
 - La création d'une machine virtuelle sous VirtualBox.
 - Le téléchargement des composants Hadoop/Spark et la configuration des chemins système (JAVA_HOME, SPARK_HOME).
 - L'exécution de scripts de traitement de données (recherche d'amis communs) pour tester le cluster.
 - La compilation de programmes Scala et leur soumission via spark-submit.

Étapes Du travail

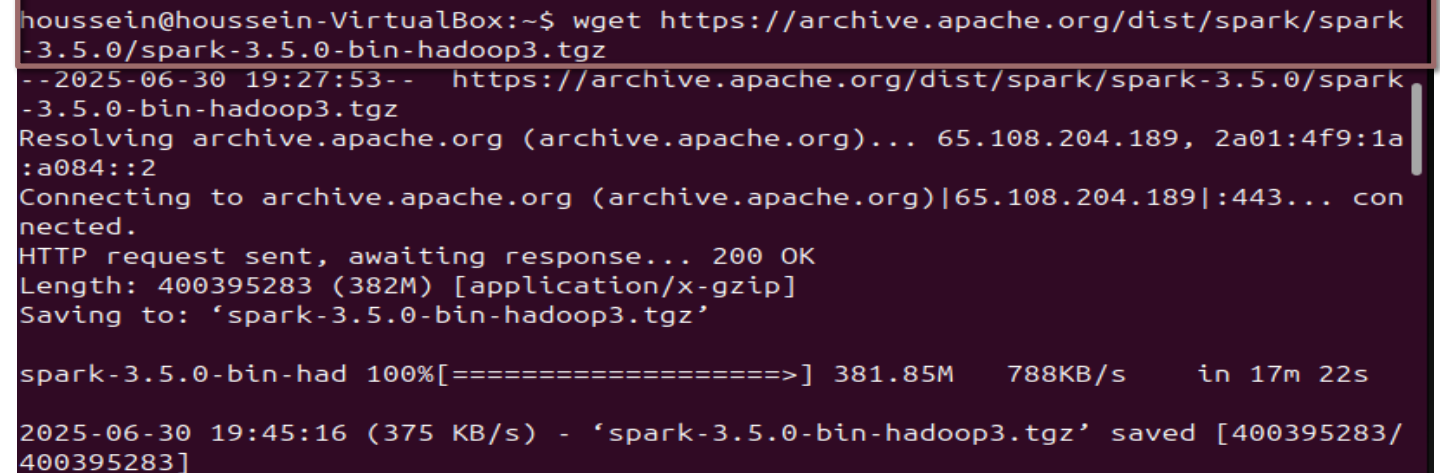
1. Installation de la Machine Virtuelle

- Installation d'Ubuntu Server sur VirtualBox.

2 .Mise en Place de la Stack Logicielle

Commandes utilisées :

- installation de spark



A terminal window showing the command to download Spark 3.5.0 using wget. The output shows the file being resolved, connected to, and saved. A red arrow points from the 'installation de spark' step to this terminal output.

```
houssein@houssein-VirtualBox:~$ wget https://archive.apache.org/dist/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz
--2025-06-30 19:27:53-- https://archive.apache.org/dist/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 400395283 (382M) [application/x-gzip]
Saving to: 'spark-3.5.0-bin-hadoop3.tgz'

spark-3.5.0-bin-had 100%[=====] 381.85M 788KB/s in 17m 22s

2025-06-30 19:45:16 (375 KB/s) - 'spark-3.5.0-bin-hadoop3.tgz' saved [400395283/400395283]
```

Étapes Du travail : implementation spark

3. extact le fichie

```
houssein@houssein-VirtualBox:~$ tar xzf spark-3.5.0-bin-hadoop3.tgz
houssein@houssein-VirtualBox:~$ ls
Backups                Public
Desktop                publickey
Documents              server_private.key
Downloads              server_public.key
hadoop-3.2.1.tar.gz    snap
Music                  spark-3.5.0-bin-hadoop3
mysql-connector-j-8.3.0 spark-3.5.0-bin-hadoop3.tgz
mysql-connector-j-8.3.0.tar.gz Templates
Pictures               tp-1.0.0.jar
poeme.txt              Videos
privatekey
```

Étapes Du travail : implementation spark

Configuration : Variables dans ~/.bashrc

```
GNU nano 7.2 /home/houssein/.bashrc
export HADOOP_HOME=/usr/local/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export JAVA_HOME=/usr/lib/jvm/java-17-openjdk-amd64
export JAVA_HOME=/usr/lib/jvm/java-17-openjdk-amd64
export SPARK_HOME=/home/houssein/spark-3.5.0-bin-hadoop3
export PATH=$PATH:$SPARK_HOME/bin
export JAVA_HOME=/usr/lib/jvm/java-17-openjdk-amd64
export SPARK_HOME=/home/houssein/spark-3.5.0-bin-hadoop3
export PATH=$PATH:$SPARK_HOME/bin
```

Étapes Du travail : implementation spark

charger spark

→ `housein@housein-VirtualBox:~$ source ~/.bashrc`

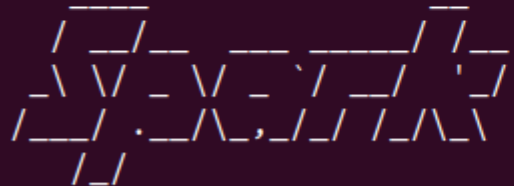
Vérification : spark-shell

`housein@housein-VirtualBox:~$ spark-shell --version`

25/06/30 20:36:46 WARN Utils: Your hostname, houssein-VirtualBox resolves to a loopback address: 127.0.1.1; using 192.168.206.135 instead (on interface enp0s3)

25/06/30 20:36:46 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address

Welcome to

 version 3.5.0

Using Scala version 2.12.18, OpenJDK 64-Bit Server VM, 17.0.15

Étapes Du travail : implementation spark

charger spark

→ `housein@housein-VirtualBox:~$ source ~/.bashrc`

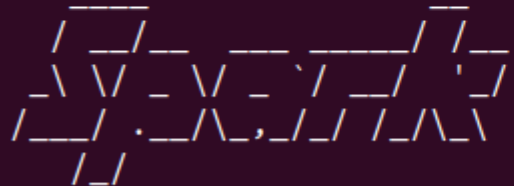
Vérification : spark-shell

`housein@housein-VirtualBox:~$ spark-shell --version`

25/06/30 20:36:46 WARN Utils: Your hostname, houssein-VirtualBox resolves to a loopback address: 127.0.1.1; using 192.168.206.135 instead (on interface enp0s3)

25/06/30 20:36:46 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address

Welcome to

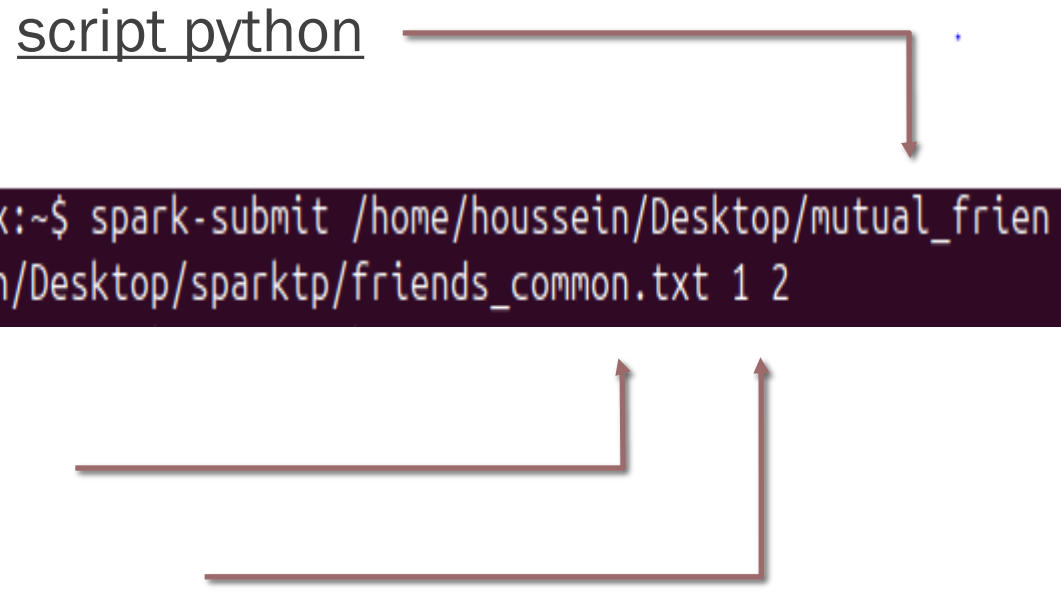
 version 3.5.0

Using Scala version 2.12.18, OpenJDK 64-Bit Server VM, 17.0.15

Étapes Du travail :implementation pyspark

Pip install pyspark

script python



```
housein@housein-VirtualBox:~$ spark-submit /home/houssein/Desktop/mutual_friends_pyspark.py /home/houssein/Desktop/sparktp/friends_common.txt 1 2
```

The diagram illustrates the components of the terminal command. A red arrow points from the text 'script python' to the file path '/home/houssein/Desktop/mutual_friends_pyspark.py' in the command. Two red arrows point from the text 'File test' and 'ids' to the file path '/home/houssein/Desktop/sparktp/friends_common.txt' in the command.

File test

ids

Étapes Du travail :implementation pyspark

Exemple de test

```
houssein@houssein-VirtualBox:~$ cat output/mutual_friends/part-00000  
1Sidi2Mohamed 3
```

Étapes Du travail :implementation Scala

Install scala →

```
houssein@houssein-VirtualBox:~$ scala -version
Scala code runner version 2.12.18 -- Copyright 2002-2023, LAMP/EPFL and
Lightbend, Inc.
houssein@houssein-VirtualBox:~$ java -version
openjdk version "17.0.15" 2025-04-15
```

Compiler le programme en scala

```
houssein@houssein-VirtualBox:~$ scalac -classpath "$SPARK_HOME/jars/*"
/home/houssein/Desktop/mutual_friends.scala
```

Créer un fichier JAR →

```
houssein@houssein-VirtualBox:~$ jar -cvf MutualFriends.jar *.class
added manifest
```

Étapes Du travail :implementation Scala

Exécuter avec spark-submit

```
houssein@houssein-VirtualBox:~$ spark-submit --class MutualFriends Mutu  
alFriends.jar
```

Exemple de test



```
houssein@houssein-VirtualBox:~$ cat output/part-00000
0      3      4
3      4      0
1      2      0,4
0      1      2,4
houssein@houssein-VirtualBox:~$ cat output/part-00001
0      2      1,4
1      4      0,2
2      4      0,1
0      4      1,2,3
houssein@houssein-VirtualBox:~$ cat output1/part-00001
0      4      1,2,3
1      2      0,4
```

Demonstration
