

Installer Spark en architecture distribué

INSTALLER SPARK:

(Spark est un outil de traitement distribué permettant de faire des map/reduce et de travailler en parallèle sur plusieurs machines)

Télécharger Spark: <https://spark.apache.org/downloads.html>

Installer Spark sur le windows : C:\Program Files\spark-3.5.5-bin-hadoop3

Installer de dossier "hadoop"(zip) a la racine du windows:
C:\hadoop\

Télécharger le JDK de java:(java version 19.0.2 pour spark version 3-5-5)
<https://www.oracle.com/java/technologies/javase/jdk19-archive-downloads.html>

Installer le JDK (en lançant l'exécutable)

Installer le java sur C:\Program Files\Java\jdk-19

Tester le java sur en terminal:

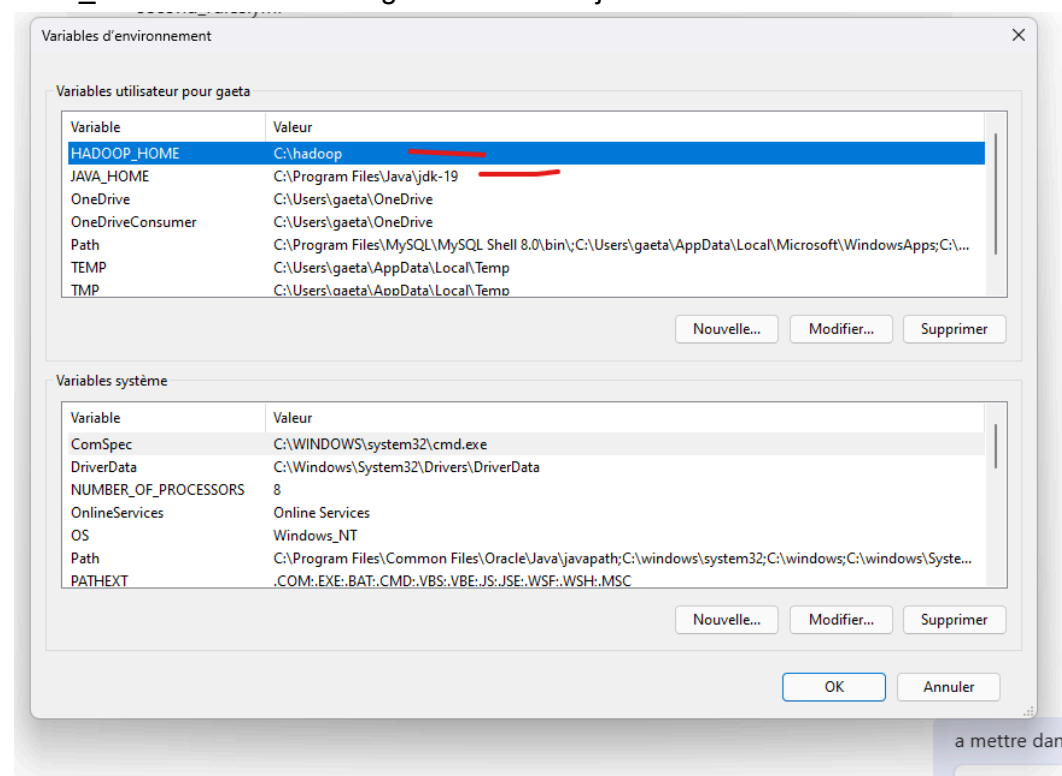
java -version

créer 2 variables d'environnement sur les variables d'environnement systèmes de windows (respecter le nommage des variables):

- Le premier a l'url du folder hadoop
- le second a l'url du java

HADOOP_HOME C:\hadoop

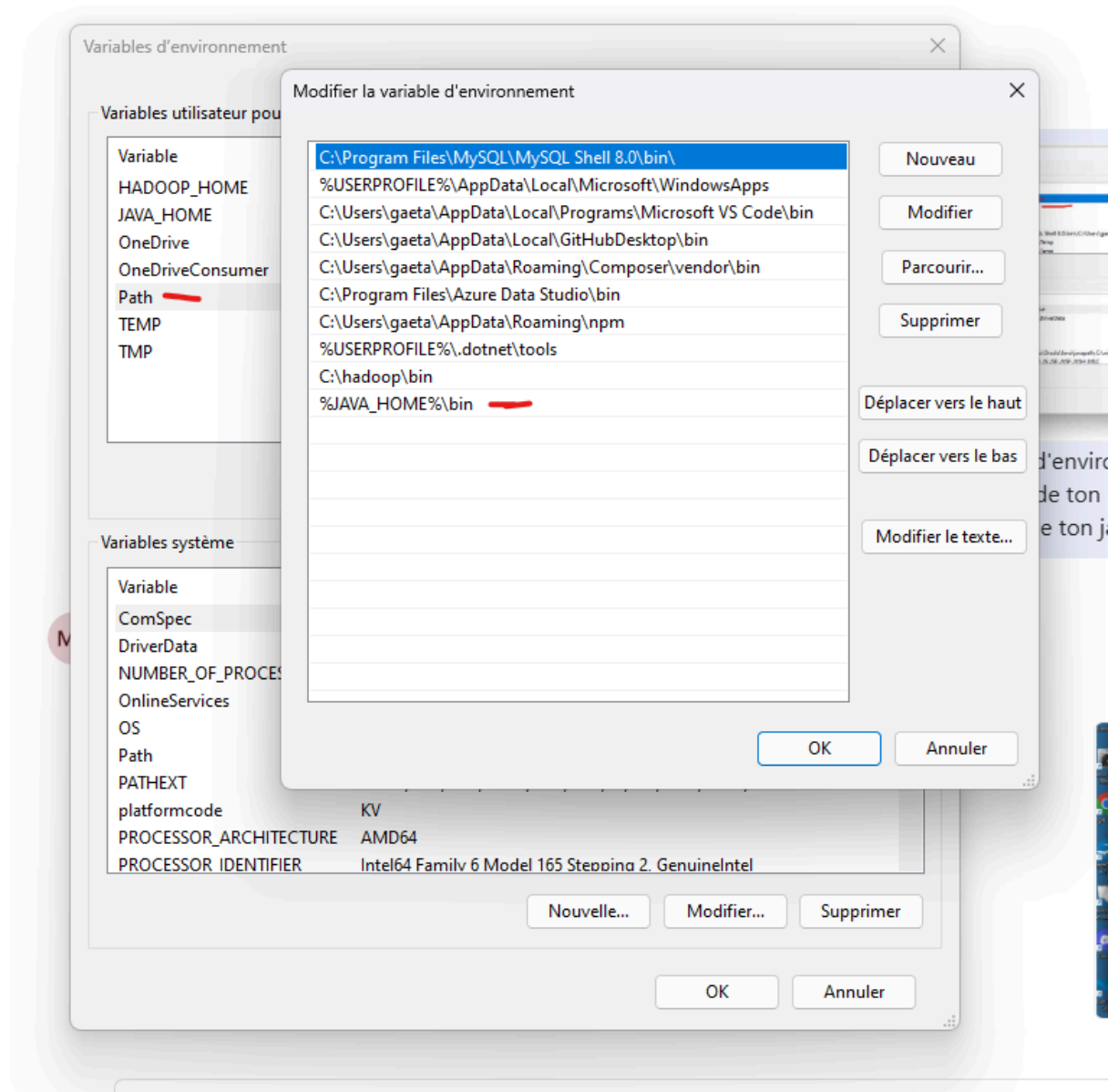
JAVA_HOME C:\Program Files\Java\jdk-19



créer 1 variables d'environnement sur les variables d'environnement systèmes de windows (respecter le nommage des variables):

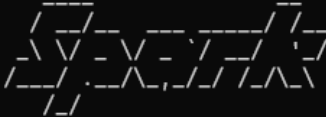
- dans le Path, indiquer comment windows doit utiliser java

%JAVA_HOME%\bin



Aller dans le programme de Spark (C:\Program Files\spark-3.5.5-bin-hadoop3)
 et ouvrir un terminal:
 ./bin/spark-shell

```
PS C:\Program Files\spark-3.5.5-bin-hadoop3> ./bin/spark-shell  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
Spark context Web UI available at http://LAPTOP-TPP15DD7:4040  
Spark context available as 'sc' (master = local[*], app id = local-1745823027463).  
Spark session available as 'spark'.  
Welcome to
```



```
version 3.5.5  
  
Using Scala version 2.12.18 (Java HotSpot(TM) 64-Bit Server VM, Java 19.0.2)  
Type in expressions to have them evaluated.  
Type :help for more information.
```

```
scala>
```

Aller sur le <http://localhost:4040/>

Connexion a Spark en fonctionnement réussi

INSTALLER ZOOKEPPER:

(zookeeper est un orchestrateur de Spark.

C'est lui qui defini le noeud maitre et les noeud esclave de spark. Il est normal d'avoir plusieurs zookeeper pour que si un zookeeper s'arrete, un autre puisse prendre le relais)

télécharger zookeeper:

(il faut un ZooKeeper 3.8.x ou 3.9.x (compatible Hadoop 3, Java 8+))

(il faut la version classique, pas la "source release").

<https://zookeeper.apache.org/releases.html#download>

Dézipper, et installer la folder de zookeeper dans:

C:\apache-zookeeper-3.9.3-bin

FAIRE COMMUNIQUER ZOOKEEPER ET SPARK:

POUR ZOOKEEPER:

voir son adresse IP:

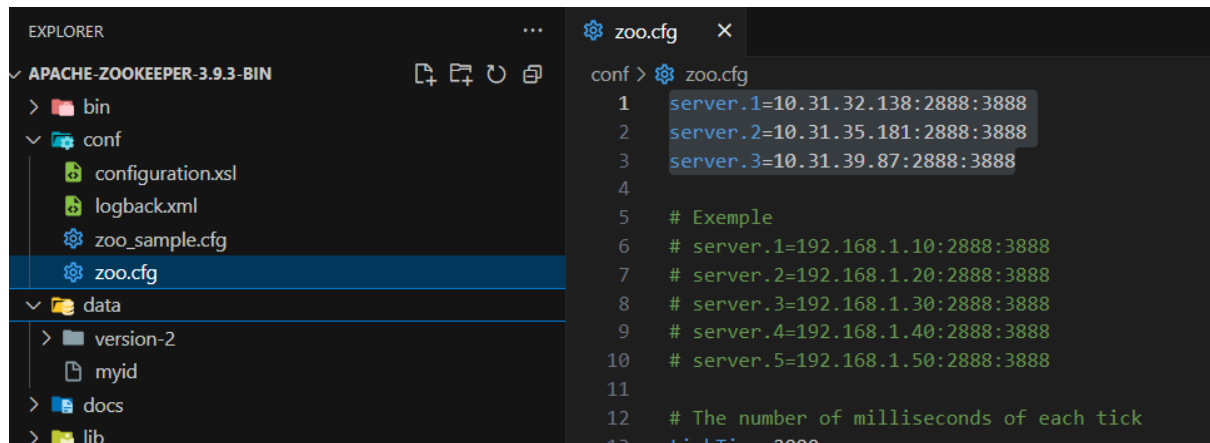
sur le terminal faire ipconfig et prendre le IPV4: (ex 10.31.35.181)

Dans zookeeper, dans le fichier conf/zoo.cfg, définir l'adresse IP de chaque ZOOKEEPER (2888:3888 sont les ports par défaut le crois...)

server.1=10.31.32.138:2888:3888

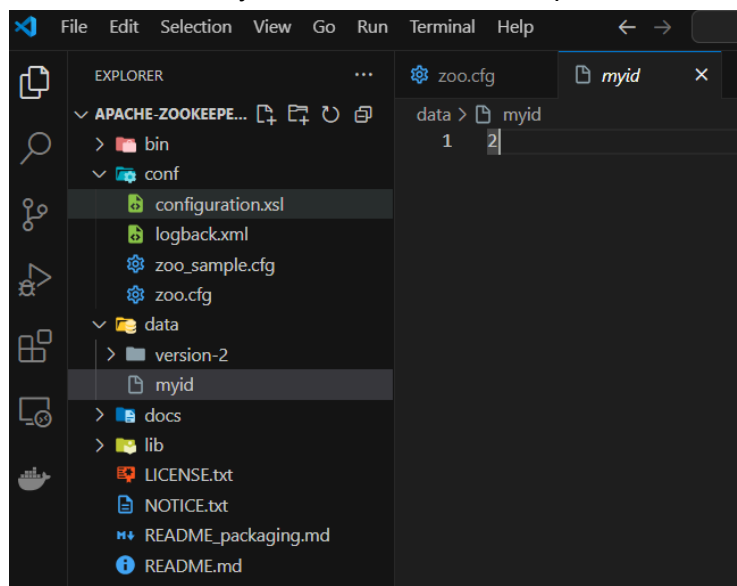
server.2=10.31.35.181:2888:3888

server.3=10.31.39.87:2888:3888



Dans zookeeper, créer un dossier data

Créer un fichier myid avec le chiffre correspondant un numéro du serveur correspondant

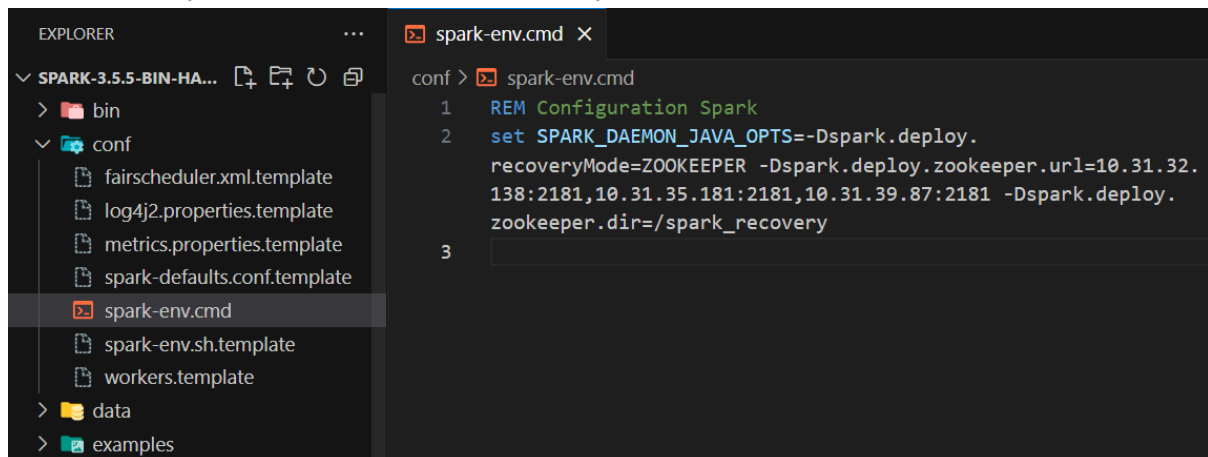


POUR SPARK:

dans le fichier conf/spark-env.cmd (se baser sur spark-default.conf.template),
ajouter les paramètres pour paramétrer spark en lui définissant les zookeeper qui vont le
manager. (effacer tout les commentaires)

REM Configuration Spark

```
set SPARK_DAEMON_JAVA_OPTS=-Dspark.deploy.recoveryMode=ZOOKEEPER  
-Dspark.deploy.zookeeper.url=10.31.32.138:2181,10.31.35.181:2181,10.31.39.87:2181  
-Dspark.deploy.zookeeper.dir=/spark_recovery
```



Tester les ip:

pour chaque machine, récupérer les ip avec ipconfig
essayer de se pinger entre machine
ping adresse_ip

si cela ne marche pas, c'est probablement le parefeu windows.
ouvrir sur la machine qui ne repond pas sur le terminal administrateur:
netsh advfirewall set allprofiles state off

Lancer SPARK ET ZOOKEEPER:

- relancer Spark MASTER:

aller dans C:\Program Files\spark-3.5.5-bin-hadoop3\bin

ouvrir le terminal:

(definir le IP d'instance spark, et un PORT webui différent pour chaque instance)

```
.\spark-class.cmd org.apache.spark.deploy.master.Master --host 10.31.35.181 --port 7077  
--webui-port 8081
```

- lancer Zookeeper:

aller dans C:\apache-zookeeper-3.9.3-bin\bin

ouvrir le terminal:

```
./zkServer.cmd
```

- Lancer Spark WORKER:

Sur le terminal Powershell en administrateur:

aller dans C:\Program Files\spark-3.5.5-bin-hadoop3\bin

ouvrir le terminal:

```
.\spark-class.cmd org.apache.spark.deploy.worker.Worker
```

```
spark://10.31.32.138:7077,10.31.35.181:7077,10.31.39.87:7077
```