

Configuration Zookeeper, Spark, Zeppelin

Configuration Zookeeper	1
1. Installation	1
2. Configuration	2
2.1 Fichier zoo.cfg	2
Configuration Spark	2
3. Installation	3
4. Configuration	3
4.1 Configuration de la haute disponibilité dans Spark	3
Configuration Zeppelin	4
5. Installation	4
6. Configuration	4
Démarrage Processus	5
7. Démarrage Zookeeper	5
8. Démarrage Spark	5
8.1 Démarrage Masters	5
8.2 Démarrage Slaves	5
9. Démarrage Zeppelin	7

Configuration Zookeeper

1. Installation

Télécharger la version 3.4.12 (zookeeper-3.4.12.tar.gz) depuis l'adresse suivante :
wget <https://archive.apache.org/dist/zookeeper/zookeeper-3.4.12/zookeeper-3.4.12.tar.gz>

Sur les trois machines où le spark masters vont être installés suivre les étapes suivantes.

Décompresser l'archive dans un répertoire de preference \$HOME

```
tar xvfz zookeeper-3.4.12.tar.gz
```

Créer les répertoires data et logs de zookeeper :

```
mkdir -p $HOME/zookeeper-3.4.12/data
```

```
mkdir -p $HOME/zookeeper-3.4.12/logs
```

Vérifier que l'on a bien la variable d'environnement JAVA_HOME, sinon la positionner avec la commande suivante :

```
export JAVA_HOME=/etc/alternatives/jre
```

2. Configuration

Sur les trois machines où le spark masters vont être installés suivre les étapes suivantes.

2.1 Fichier zoo.cfg

Aller dans le sous-répertoire conf de zookeeper et initialiser le fichier de configuration :

```
cd $HOME/zookeeper-3.4.12/conf
```

```
cp zoo_sample.cfg zoo.cfg
```

Changer le contenu des variables suivantes :

```
dataDir=/home/hadoop/zookeeper-3.4.12/data
```

Ajouter le contenu suivant à la fin du fichier:

```
dataLogDir=/home/hadoop/zookeeper-3.4.12/logs/
```

```
server.1=spark1:2888:3888
```

```
server.2=spark2:2889:3889
```

```
server.3=spark3:2890:3890
```

Copier ce fichier sur les deux autres machines avec :

```
scp zoo.cfg hadoop@spark2:/home/hadoop/zookeeper-3.4.12/conf
```

Créer un fichier « myid » dans le sous-répertoire data de zookeeper

- Sur la 1ere machine, juste mettre 1 dans ce fichier et sauvegarder
- Sur la 2eme machine, mettre 2 dans le fichier myid et sauvegarder

Configuration Spark

3. Installation

Sur chaque machine télécharger le package Spark avec la commande:

wget <http://apache.crihan.fr/dist/spark/spark-2.3.2/spark-2.3.2-bin-hadoop2.7.tgz>

Dezipper le package :

```
tar xvfz spark-2.3.2-bin-hadoop2.7.tgz
```

4. Configuration

Aller dans le répertoire conf de spark

```
cd spark-2.3.2-bin-hadoop2.7/conf/
```

Modifier le fichier spark-env.sh

```
cp spark-env.sh.template spark-env.sh
```

```
vi spark-env.sh
```

Insérer les lignes suivantes à la fin du fichier :

```
export SPARK_MASTER_PORT=7177
```

```
export SPARK_MASTER_WEBUI_PORT=8180
```

```
export SPARK_WORKER_WEBUI_PORT=8181
```

```
export JAVA_HOME=/etc/alternatives/jre
```

```
export SPARK_WORKER_MEMORY=2g
```

Pour l'ensemble de machines renseigner les valeurs suivantes :

spark1: 7177, 8180 et 8181

spark2: 7277, 8280 et 8281

spark3: 7377, 8380 et 8381

4.1 Configuration de la haute disponibilité dans Spark

Création du fichier « ha.conf » dans \$SPARK_HOME et ajouter le contenu suivant :

```
cd spark-2.3.2-bin-hadoop2.7/
```

```
vi ha.conf
```

```
spark.deploy.recoveryMode=ZOOKEEPER
```

```
spark.deploy.zookeeper.url=spark1:2181,spark2:2181,spark3:2181
```

```
spark.deploy.zookeeper.dir=/home/hadoop/spark-2.3.2-bin-hadoop2.7/spark
```

Créer le répertoire suivant dans spark :

```
mkdir -p /home/hadoop/spark-2.3.2-bin-hadoop2.7/spark
```

Copier le même fichier sur les deux autres machines

```
scp ha.conf hadoop@spark2:/home/hadoop/spark-2.3.2-bin-hadoop2.7/
```

Configuration Zeppelin

5. Installation

Télécharger la version zeppelin-0.8.0:

```
wget https://www-eu.apache.org/dist/zeppelin/zeppelin-0.8.0/zeppelin-0.8.0-bin-all.tgz
```

Décompresser le package:

```
tar xvfz zeppelin-0.8.0-bin-all.tgz
```

6. Configuration

Editer le fichier zeppelin-site.xml

```
cd /home/hadoop/zeppelin-0.8.0-bin-all/conf/
```

```
cp zeppelin-site.xml.template zeppelin-site.xml
```

Modifier le port d'accès de zeppelin (8090) :

```
vi zeppelin-site.xml
```

```
<property>
<name>zeppelin.server.port</name>
<value>8090</value>
<description>Server port.</description>
</property>
```

Modifier le fichier zeppelin-env.sh

```
cp zeppelin-env.sh.template zeppelin-env.sh
```

```
vi zeppelin-env.sh
```

Insérer les lignes suivantes à la fin du fichier :

```
export JAVA_HOME=/etc/alternatives/jre
export SPARK_HOME=/home/hadoop/spark-2.3.2-bin-hadoop2.7
```

Démarrage Processus

7. Démarrage Zookeeper

Démarrer le process zookeeper :

```
cd $HOME/zookeeper-3.4.12/
bin/zkServer.sh start
```

Vérifier que le processus est bien démarré :

```
jps
13200 StatePusher
5349 Main
29974 Jps
29871 QuorumPeerMain
```

Sur les trois machines vérifier le mode de fonctionnement de chaque processus Zookeeper :

```
bin/zkServer.sh status
ZooKeeper JMX enabled by default
Using config: /home/hadoop/zookeeper-3.4.12/bin/../conf/zoo.cfg
Mode: follower
```

```
bin/zkServer.sh status
ZooKeeper JMX enabled by default
Using config: /home/hadoop/zookeeper-3.4.12/bin/../conf/zoo.cfg
Mode: leader
```

8. Démarrage Spark

8.1 Démarrage Masters

Démarrer les trois masters de spark avec la commande suivante:

```
sbin/start-master.sh --properties-file ha.conf
```

Vérifier le démarrage avec la commande jps

8.2 Démarrage Slaves

Démarrer les trois slaves de spark avec les commandes suivantes sur chaque machine:

spark1:

sbin/start-slave.sh spark1:7177

spark2:

sbin/start-slave.sh spark2:7277

spark3:

sbin/start-slave.sh spark3:7377

Vérifier le démarrage avec la commande jps :

13200 StatePusher

24833 Worker

14274 Master

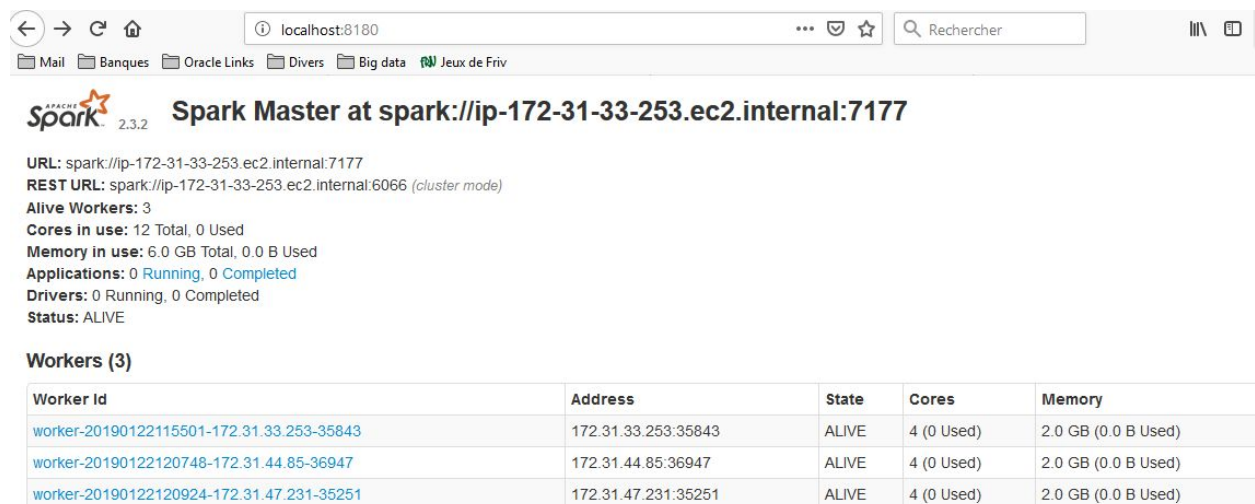
5349 Main

24919 Jps

29871 QuorumPeerMain

Vérification avec l'interface web :

Sur spark1 (master actif):



Spark Master at spark://ip-172-31-33-253.ec2.internal:7177

URL: spark://ip-172-31-33-253.ec2.internal:7177
REST URL: spark://ip-172-31-33-253.ec2.internal:6066 (cluster mode)

Alive Workers: 3
Cores in use: 12 Total, 0 Used
Memory in use: 6.0 GB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (3)

Worker Id	Address	State	Cores	Memory
worker-20190122115501-172.31.33.253-35843	172.31.33.253:35843	ALIVE	4 (0 Used)	2.0 GB (0.0 B Used)
worker-20190122120748-172.31.44.85-36947	172.31.44.85:36947	ALIVE	4 (0 Used)	2.0 GB (0.0 B Used)
worker-20190122120924-172.31.47.231-35251	172.31.47.231:35251	ALIVE	4 (0 Used)	2.0 GB (0.0 B Used)

Sur spark2 (master pasif) :

