Rapport de TP-1 : Installation et Configuration de Hadoop

Réalisé par : LAABID ABDESSAMAD

Github : https://github.com/aplusInDev/hadoop_tps/tree/main/tp1

Introduction

Ce rapport présente l'installation et la configuration de Hadoop 3.4.0 sur une machine virtuelle Ubuntu 24.04.2 utilisant VM Virtual Box. L'objectif principal est de mettre en place un environnement Hadoop fonctionnel pour le traitement distribué des données.

Environnement de travail

- **Machine virtuelle** : VM Virtual Box
- **Système d'exploitation** : Ubuntu Server 24.04.2
- **Version de Hadoop** : 3.4.0
- **Version de Java** : OpenJDK 11

Processus d'installation

1. Installation des prérequis

Installation de Java

Hadoop étant écrit en Java, l'installation d'un JDK est nécessaire. Pour ce TP, nous avons utilisé OpenJDK 11 :

```
sudo apt update
sudo apt install -y openjdk-11-jdk
```

Vérification de l'installation Java :

```
java -version
javac -version
```

Configuration SSH

Hadoop nécessite un accès SSH pour gérer les nœuds (même en configuration single-node) :

```
sudo apt install openssh-server openssh-client -y
```

Création des clés SSH pour permettre la connexion sans mot de passe :

```
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

2. Installation de Hadoop

Téléchargement et extraction de Hadoop 3.4.0 :

```
wget https://dlcdn.apache.org/hadoop/common/hadoop-3.4.0/hadoop-3.4.0.tar.gz
tar xzf hadoop-3.4.0.tar.gz
```

3. Configuration de l'environnement

Ajout des variables d'environnement Hadoop dans le fichier `.bashrc` :

```
#Hadoop Related Options
export HADOOP_HOME=/home/user/hadoop-3.4.0
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:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
```

Définition de la variable JAVA_HOME dans `hadoop-env.sh` :

```
echo "export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64" >>
$HADOOP_HOME/etc/hadoop/hadoop-env.sh
```

Configuration des fichiers XML

1. core-site.xml

Ce fichier contient la configuration générale de la plateforme Hadoop :

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
```

```
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/home/user/tmpdata</value>
  </property>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://0.0.0.0:9000</value>
  </property>
</configuration>
```

Ce fichier définit :

- Le répertoire temporaire pour Hadoop
- Le système de fichiers par défaut (HDFS) et son point d'accès sur le port 9000

2. hdfs-site.xml

Ce fichier configure le système de fichiers HDFS :

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/home/user/dfsdata/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/home/user/dfsdata/datanode</value>
  </property>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.http-address</name>
    <value>0.0.0.0:9870</value>
```

```
  </property>
</configuration>
```

Ce fichier spécifie :

- Les emplacements de stockage pour le NameNode et le DataNode
- Le facteur de réplication (1 pour un single-node cluster)
- L'adresse HTTP du NameNode (accessible via le port 9870)

3. mapred-site.xml

Ce fichier contient les paramètres spécifiques à MapReduce :

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
```

Cette configuration indique que MapReduce utilisera YARN comme framework d'exécution.

4. yarn-site.xml

Ce fichier configure le gestionnaire de ressources YARN :

```
<?xml version="1.0"?>
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
  <property>
```

```
    <name>yarn.resourcemanager.hostname</name>
    <value>0.0.0.0</value>
  </property>
  <property>
    <name>yarn.resourcemanager.webapp.address</name>
    <value>0.0.0.0:8088</value>
  </property>
  <property>
    <name>yarn.acl.enable</name>
    <value>0</value>
  </property>
  <property>
    <name>yarn.nodemanager.env-whitelist</name>

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH
_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
  </property>
</configuration>
```

Cette configuration :

- Active le service de shuffle pour MapReduce
- Définit l'adresse du ResourceManager
- Configure l'interface web du ResourceManager sur le port 8088
- Désactive le contrôle d'accès
- Spécifie les variables d'environnement autorisées

Initialisation et démarrage de Hadoop

Formatage du NameNode

Avant de démarrer les services Hadoop, le système de fichiers HDFS doit être formaté :

```
hdfs namenode -format
```

Démarrage des services

Lancement des services HDFS et YARN :

```
start-dfs.sh
```

```
start-yarn.sh
```

Vérification du fonctionnement

Pour s'assurer que tous les services sont correctement démarrés, nous utilisons la commande `jps` qui liste les processus Java en cours d'exécution :

```
user@hadoop:~$ jps
12292 Jps
11862 NodeManager
11527 SecondaryNameNode
11292 DataNode
11150 NameNode
11726 ResourceManager
user@hadoop:~$
```

Les processus suivants doivent être présents :

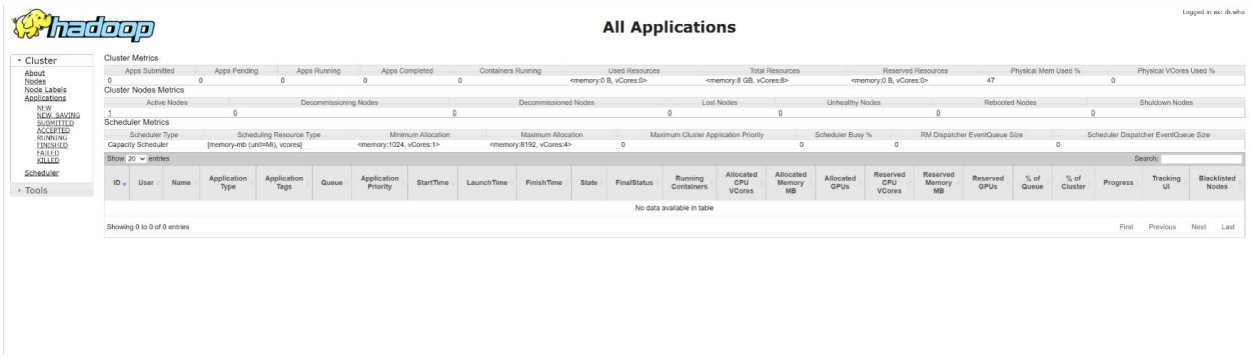
- NameNode
- DataNode
- SecondaryNameNode
- ResourceManager
- NodeManager

Accès aux interfaces web

Pour vérifier le bon fonctionnement de l'installation, nous pouvons accéder aux interfaces web :

```
user@hadoop:~$ ip a
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host noprefixroute
       valid_lft forever preferred_lft forever
2: enp0s3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc fq_codel state UP group default qlen 1000
    link/ether 08:00:27:a8:10:0e brd ff:ff:ff:ff:ff:ff
    inet 192.168.11.105/24 metric 100 brd 192.168.11.255 scope global dynamic enp0s3
       valid_lft 80292sec preferred_lft 80292sec
    inet6 fe80::a00:27ff:fea8:100e/64 scope link
       valid_lft forever preferred_lft forever
```

- Interface du ResourceManager : http://192.168.11.105:8088/



- Interface du NameNode : http://192.168.11.105:9870/

Hadoop | Overview | Datanodes | Datanode Volume Failures | Snapshot | Startup Progress | Utilities

Overview '0.0.0.0:9000' (✔active)

Started:	Thu May 01 08:10:56 +0100 2025
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled:	Mon Mar 04 07:35:00 +0100 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID-910caddb-1873-4f9e-a352-888a8e7adbaa
Block Pool ID:	BP-259729719-127.0.1.1-1746083429441

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 77.14 MB of 169 MB Heap Memory. Max Heap Memory is 980 MB.

Non Heap Memory used 50.94 MB of 54.44 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	19.52 GB
Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)
Non DFS Used:	9.48 GB
DFS Remaining:	9.02 GB (46.24%)
Block Pool Used:	24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Thu May 01 08:10:56 +0100 2025
Last Checkpoint Time	Thu May 01 08:10:29 +0100 2025
Last HA Transition Time	Never
Enabled Erasure Coding Policies	RS-6-3-1024k

NameNode Journal Status

Current transaction ID: 1

Journal Manager	State
FileJournalManager(root=/home/user/dfsdata/namenode)	EditLogFileOutputStream(/home/user/dfsdata/namenode/current/edits_inprogress_0000000000000000001)

NameNode Storage

Storage Directory	Type	State
/home/user/dfsdata/namenode	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	19.52 GB	24 KB (0%)	9.02 GB (46.24%)	24 KB	1

Conclusion

Ce TP nous a permis d'installer et de configurer avec succès un environnement Hadoop 3.4.0 sur une machine virtuelle Ubuntu 24.04.2. Cette installation servira de base pour les prochains travaux pratiques sur le traitement distribué des données.

L'environnement mis en place est opérationnel pour exécuter des jobs MapReduce et stocker des données dans HDFS. Les configurations réalisées correspondent à un cluster single-node, qui peut être étendu pour former un cluster multi-nœuds dans de futurs travaux.