Hadoop: Installation et Configuration

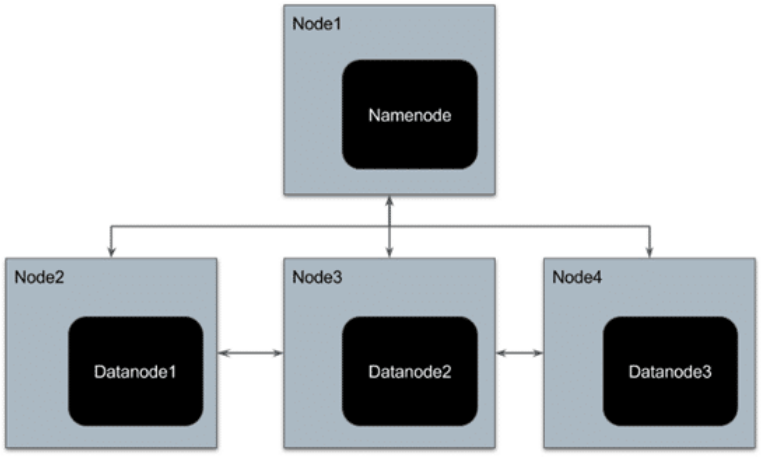
1. Présentation

Hadoop est un Framework libre qui offre un espace de stockage massif pour tous les types de données, une immense puissance de traitement et la possibilité de prendre en charge une quantité de tâches virtuellement illimitée. Basé sur Java, ce Framework fait partie du projet Apache, sponsorisé par Apache Software Foundation.

Hadoop est un composant essentiel de l'industrie du Big Data car il fournit la couche de stockage la plus fiable, HDFS, qui peut évoluer massivement. Des entreprises comme Yahoo et Facebook utilisent HDFS pour stocker leurs données.

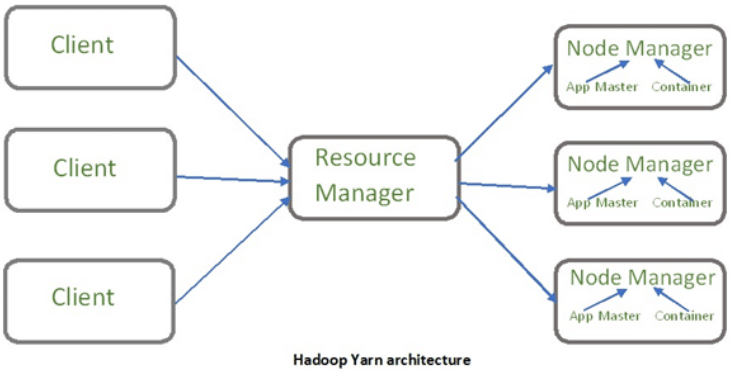
2. Overview sur HDFS

HDFS a une architecture maître-esclave où le nœud maître est appelé NameNode et le nœud esclave est appelé DataNode. Le NameNode et ses DataNodes forment un cluster. NameNode agit comme un instructeur pour DataNode tandis que les DataNodes stockent les données réelles.



source: Hasura

Il existe un autre composant de Hadoop appelé YARN. L'idée de Yarn est de gérer les ressources et de planifier/surveiller les tâches dans Hadoop. Yarn comporte deux composants principaux, Resource Manager et Node Manager. Le gestionnaire de ressources a le pouvoir d'allouer des ressources à diverses applications exécutées dans un cluster. Le gestionnaire de nœuds est chargé de surveiller leur utilisation des ressources (CPU, mémoire, disque) et d'en rendre compte au gestionnaire de ressources.



Hadoop Yarn architecture

source: GeeksforGeeks

3. Configuration

- Après avoir téléchargé Hadoop version 3.3.5, décompressez-le dans un dossier Hadoop.
- Veuillez noter lors de la création de dossiers, N'AJOUTEZ PAS D'ESPACES ENTRE LE NOM DU DOSSIER (cela peut causer des problèmes plus tard).

4. Configuration des variables d'environnement

4.1. Configuration de HADOOP_HOME

- Ouvrez la variable d'environnement et cliquez sur «Nouveau» dans «Variable utilisateur».
- Entrez le nom de la variable «HADOOP_HOME» et le chemin d'accès au dossier Hadoop comme valeur de variable. Appuyez ensuite sur « OK ».

4.2. Définition de la variable PATH

La deuxième étape de la définition de la variable d'environnement consiste à définir le chemin dans la variable PATH.

- Sélectionnez la variable PATH dans les variables système et cliquez sur «Modifier».
- Ajoutez maintenant ajouter ces chemins à la variable de chemin un par un:

1. %JAVA_HOME%\bin
2. %HADOOP_HOME%\bin
3. %HADOOP_HOME%\sbin

- Cliquez sur OK et OK. Ainsi vous avez fini avec la définition des variables d'environnement.

4.3. Vérification des chemins

Vérifiez maintenant votre configuration en ouvrant une NOUVELLE fenêtre de commande et exécutez les commandes suivantes:

```
echo %JAVA_HOME%
echo %HADOOP_HOME%
echo %PATH%
```

5. Modification des fichiers Hadoop

Une fois que vous avez configuré les variables d'environnement, l'étape suivante consiste à configurer Hadoop.

5.1. Création de dossiers

- Créez le dossier Data dans le répertoire Hadoop.
- Une fois le dossier Data créé, vous devez créer 2 nouveaux dossiers, à savoir namenode et datanode, dans le dossier Data.
- Ces dossiers sont importants car les fichiers sur HDFS résident à l'intérieur du datanode.

5.2. Modification des fichiers de configuration

Ouvrez le répertoire Hadoop -> etc -> hadoop et éditer les fichiers de configuration suivants dans hadoop pour le configurer:

```
core-site.xml
hdfs-site.xml
mapred-site.xml
yarn-site.xml
hadoop-env.cmd
```

5.2.1. Modification de core-site.xml

Faites un clic droit sur le fichier, sélectionnez Modifier et collez le contenu suivant dans les balises <configuration> </configuration>.

Note: La partie ci-dessous contient déjà la balise de configuration, vous devez copier uniquement la partie à l'intérieur.

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

5.2.2. Modification de hdfs-site.xml

Faites un clic droit sur le fichier, sélectionnez Modifier et collez le contenu suivant dans les balises <configuration></configuration>.

Remplacez également PATH~1 et PATH~2 par le chemin du dossier namenode et datanode que vous avez créés récemment (étape 5.1).

```
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>PATH~1\namenode</value>
<final>true</final>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>PATH~2\datanode</value>
<final>true</final>
</property>
</configuration>
```

5.2.3. Modification de mapred-site.xml

Faites un clic droit sur le fichier, sélectionnez Modifier et collez le contenu suivant dans les balises <configuration> </configuration>.

```
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

5.2.4. Modification de fil-site.xml

Faites un clic droit sur le fichier, sélectionnez Modifier et collez le contenu suivant dans les balises <configuration> </configuration>.

```
<configuration>
 <property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
 </property>
 <property>

<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
 </property>
<!-- Site specific YARN configuration properties --
></configuration>
```

5.2.5. Vérification de hadoop-env.cmd

Faites un clic droit sur le fichier, sélectionnez modifier et vérifiez si JAVA_HOME est défini correctement ou non. Vous pouvez remplacer la variable JAVA_HOME dans le fichier par votre JAVA_HOME réel que vous avez configuré dans la variable système:

```
set JAVA_HOME=%JAVA_HOME%
```

5.3. Remplacement du bin

La dernière étape de la configuration de Hadoop consiste à télécharger et à remplacer le dossier bin.

- Accédez au dépôt GitHub et téléchargez le dossier hadoop-3.3.5/bin sous forme de zip.
- Extrayez le zip et copiez tous les fichiers présents dans le dossier bin dans %HADOOP_HOME%\bin

6. Test

6.1. Formatting Namenode

Avant de démarrer Hadoop, vous devez formater **Namenode**. Pour cela, démarrez une NOUVELLE invite de commande et exécuter la commande ci-dessous:

```
hadoop namenode –format
```

Note: Cette commande formate toutes les données dans namenode. Il est donc conseillé de l'utiliser uniquement au début et de ne pas l'utiliser à chaque fois lors du démarrage du cluster hadoop pour éviter la perte de données.

6.2. Launching Hadoop

- Démarrez maintenant une nouvelle invite de commande.
- Changez maintenant le répertoire dans cmd en dossier sbin du répertoire hadoop, (Remarque : assurez-vous d'écrire le chemin conformément à votre système)
- Exécutez la commande ci-dessous:

```
start-all.cmd
```

Vous pouvez aussi utilisez

```
start-dfs.cmd
```

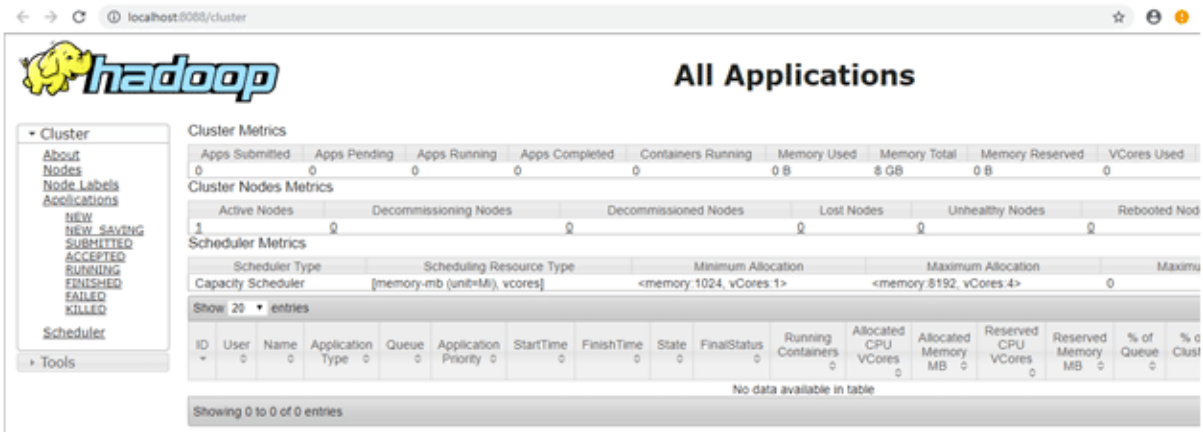
```
start-yarn.cmd
```

Cela ouvrira 4 nouvelles fenêtres cmd exécutant 4 Deamons différents de hadoop:

- Namenode
- Datanode
- Resourcemanager
- Nodemanager

6.3. Running Hadoop

- Ouvrez localhost:8088 dans un onglet de navigateur pour vérifier les détails du gestionnaire de ressources.



- Pour vérifier les détails sur le hdfs (namenode et datanode), consultez ce link sur votre navigateur.
- Pour arrêter tous les démons en cours d'exécution sur votre ordinateur, exécutez la commande ci-dessous:

```
stop-all.cmd
```

Vous pouvez aussi utilisez

```
stop-dfs.cmd
```

```
stop-yarn.cmd
```

Hadoop: Manipulation

Pour le bon déroulement de ce TP :

- Créez un répertoire TP, puis deux sous-répertoires code et data dans lesquels vous sauvegarderai respectivement les codes de vos mappers et reducers, et les données sources et résultat.
- Déplacez-vous sous le répertoire ~/TP/data, et y importer le fichier purchases.txt.

Toutes les commandes interagissant avec le système Hadoop commencent par hadoop fs. Ensuite, les options rajoutées sont très largement inspirées des commandes Unix standard.

1. HDFS

1. Créer un répertoire dans HDFS, appelé input. Pour cela, taper:

```
hadoop fs –mkdir input
```

2. Pour copier le fichier purchases.txt dans HDFS sous le répertoire input, taper la commande:

```
hadoop fs –put ~/purchases.txt input/
```

3. Pour afficher le contenu du répertoire input, la commande est:

```
hadoop fs –ls input
```

4. Pour visualiser les dernières lignes du fichier, taper:

```
hadoop fs –tail input/purchases.txt
```

Dans le tableau suivant, nous résumons les commandes les plus utilisées dans Hadoop:

hadoop fs –get file.txt	Download un fichier à partir de hadoop sur votre disque local
hadoop fs –tail file.txt	Lire les dernières lignes du fichier
hadoop fs –cat file.txt	Affiche tout le contenu du fichier
hadoop fs –mv file.txt newfile.txt	Renommer le fichier
hadoop fs –rm newfile.txt	Supprimer le fichier
hadoop fs –rm -r dossier	Supprimer le dossier
hadoop fs –mkdir myinput	Créer un répertoire
hadoop fs –cat file.txt | less	Lire le fichier page par page