Cours de « Apache Spark »

Introduction

Interface de compilation online pour Scala :

https://scastie.scala-lang.org/

Install Hadoop & Spark on Windows

1. **Download**

Hadoop 3.xxx https://hadoop.apache.org/releases.html

Java 8.xxx https://adoptopenjdk.net/

Spark https://spark.apache.org/downloads.html

Download https://www.7-zip.org/

Hadoop winutils https://github.com/steveloughran/winutils

2. **Optional downloads for working with python**

PyThon https://www.python.org/downloads/windows/

PySpark https://pypi.org/project/pyspark/

3. **Créer les répertoires suivants :**

C:\Hadoop

C:\Spark

C:\Java

C:\Hadoop\datanode

C:\Hadoop\namenode

4. **Mettre chaque élément dans son emplacement**

   Décompresser hadoop et le déplacer dans C:\Hadoop
   Décompresser Spark et le déplacer dans C:\Spark
   Installer (ou décompresser java) et le déplacer dans C:\Java
   Décompresser winutils et copier le contenu du ..hadoop3..\bin dans c:\hadoop\bin

5. **Définir les variables d'environnement**

Ajouter la variable d'environnement JAVA_HOME= C:\Java

Ajouter la variable d'environnement HADOOP_HOME= C:\Hadoop

Ajouter la variable d'environnement SPARK_HOME= C:\Spark

Modifier la variable d'enviromment Path et rajouter :

%JAVA_HOME%\bin

%HADOOP_HOME%\bin

%HADOOP_HOME%\sbin

%SPARK_HOME%\bin

Echo %PATH% pour vérifier

A ce niveau vous pouvez tester par :

Java -version

Hdfs -version

Spark-shell --version

6. Configuration hadoop

Next, go to %HADOOP_HOME%\etc\hadoop and edit (or create) the file core-site.xml so it looks like the following:

6.1 core-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9900</value>
  </property>
</configuration>
```

In the same directory, edit (or create) mapred-site.xml with the following contents:

6.2 mapred-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

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

Next, edit (or create) hdfs-site.xml:

6.3 hdfs-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
 <property>
  <name>dfs.replication</name>
  <value>1</value>
 </property>
 <property>
  <name>dfs.namenode.name.dir</name>
  <value>file:///C:/Hadoop/hadoop/namenode</value>
 </property>
 <property>
  <name>dfs.datanode.data.dir</name>
  <value>file:///C:/Hadoop/hadoop/datanode</value>
 </property>
</configuration>
```

Finally, edit yarn-site.xml so it reads:

6.5 yarn-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<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>
</configuration>
```

6.6 File access

winutils chmod 777 c:\Hadoop\datanode

winutils chmod 777 c:\Hadoop\namenode

7. Lancer l'environnement hadoop

Start-dfs ➔ pour lancer hadoop
Start-yarn ➔ pour lancer yarn

Ou

start-all ➔ pour lancer l'ensemble

8. Lancer spark

spark-shell

9. Vérification

La commande jps

Le résultat sera comme suit :

c:\Hadoop>jps

18160 NodeManager

3456 SparkSubmit

16488 NameNode

6520 Jps

17980 DataNode

7964 SourceManager

10. Interfaces UI Nécessaires

http://127.0.0.1:8088/ ➔ Hadoop Yarn
http://127.0.0.1:9870/ ➔ Hadoop Datanodes
http://127.0.0.1:4040/ ➔ Spark UI