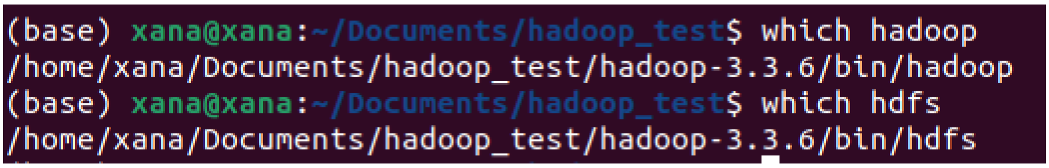
Extraction des données depuis un système Big Data

0. Import des données vers HDFS

On suppose ici que hadoop est installé et fonctionnel.

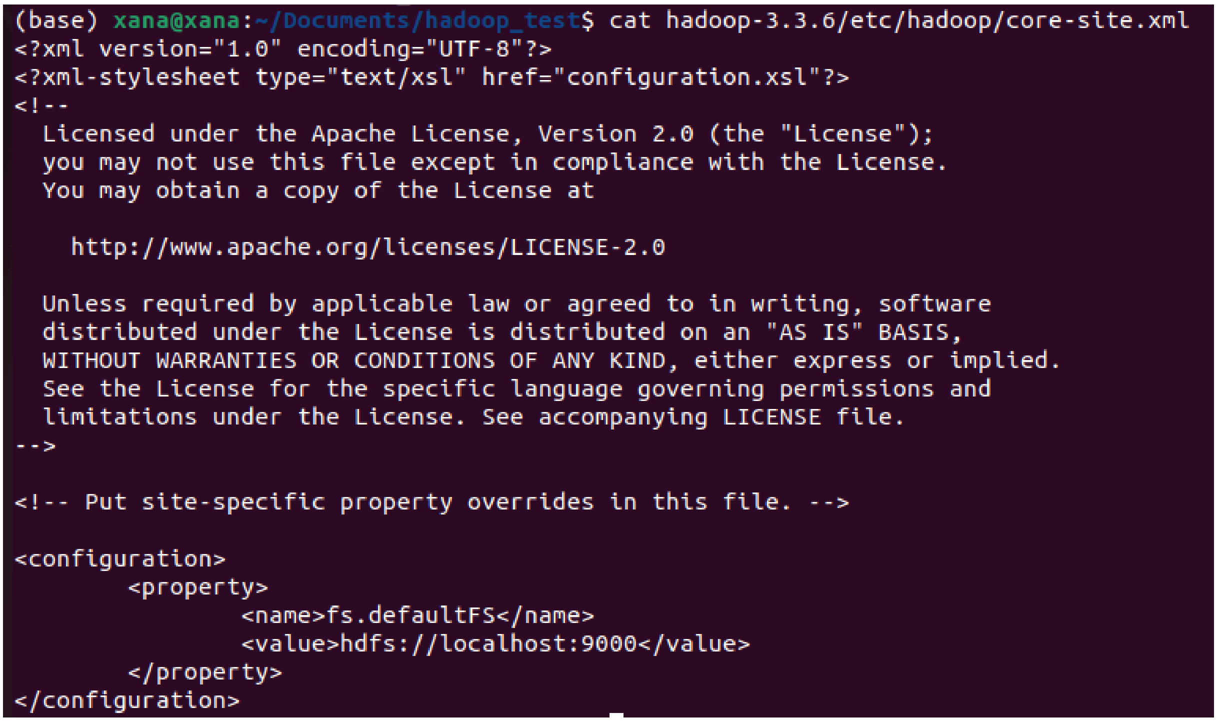


```
(base) xana@xana:~/Documents/hadoop_test$ which hadoop
/home/xana/Documents/hadoop_test/hadoop-3.3.6/bin/hadoop
(base) xana@xana:~/Documents/hadoop_test$ which hdfs
/home/xana/Documents/hadoop_test/hadoop-3.3.6/bin/hdfs
```

1. Création du Cluster (1 node)

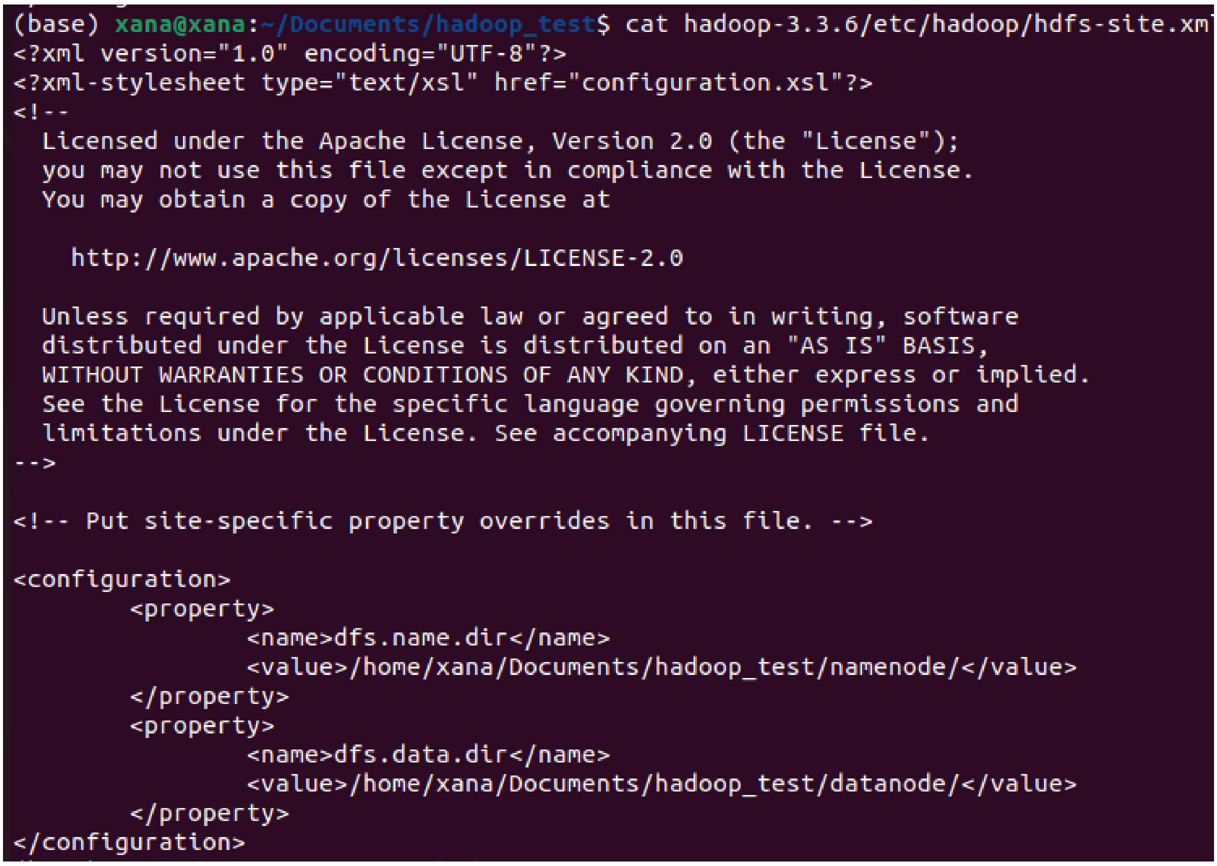
Dans un premier temps, on va créer un cluster constitué de un datanode :

0. Créer 2 dossiers datanode et namenode
1. Edition de core-site.xml



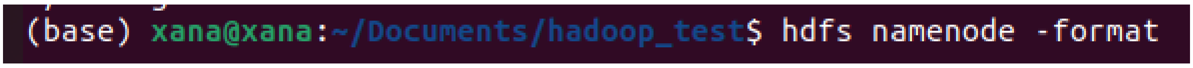
```
(base) xana@xana:~/Documents/hadoop_test$ cat hadoop-3.3.6/etc/hadoop/core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

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

2. Edition de hdfs-site.xml



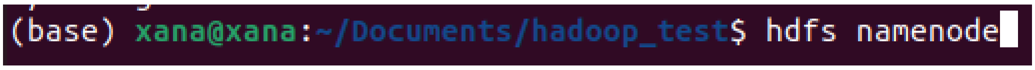
```
(base) xana@xana:~/Documents/hadoop_test$ cat hadoop-3.3.6/etc/hadoop/hdfs-site.xm
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
        <property>
                <name>dfs.name.dir</name>
                <value>/home/xana/Documents/hadoop_test/namenode/</value>
        </property>
        <property>
                <name>dfs.data.dir</name>
                <value>/home/xana/Documents/hadoop_test/datanode/</value>
        </property>
</configuration>
```

3. Formatage du système de fichier :

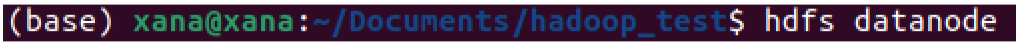


```
(base) xana@xana:~/Documents/hadoop_test$ hdfs namenode -format
```

4. Vérification : Dans 2 autres terminals lancer :



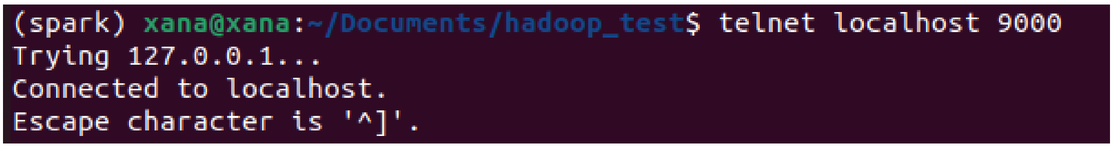
```
(base) xana@xana:~/Documents/hadoop_test$ hdfs namenode
```



```
(base) xana@xana:~/Documents/hadoop_test$ hdfs datanode
```

Repérer les ports attribués et vérifier si la connexion peut être établie sur namenode et datanode :

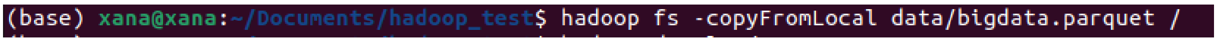
telnet localhost <portnumber>
Démonstration de la réussite :



```
(spark) xana@xana:~/Documents/hadoop_test$ telnet localhost 9000
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
```

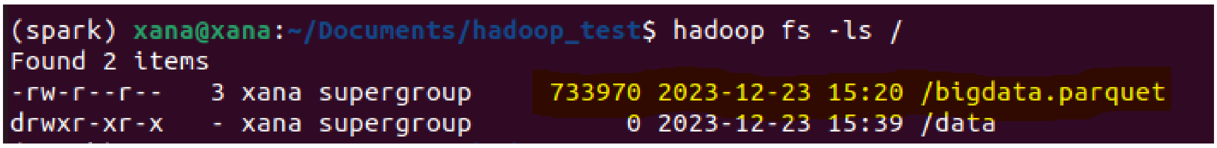
2. Import des données

Dans le cadre d'une réalisation fictive , on va importer un .parquet dans hdfs par la commande :



```
(base) xana@xana:~/Documents/hadoop_test$ hadoop fs -copyFromLocal data/bigdata.parquet /
```

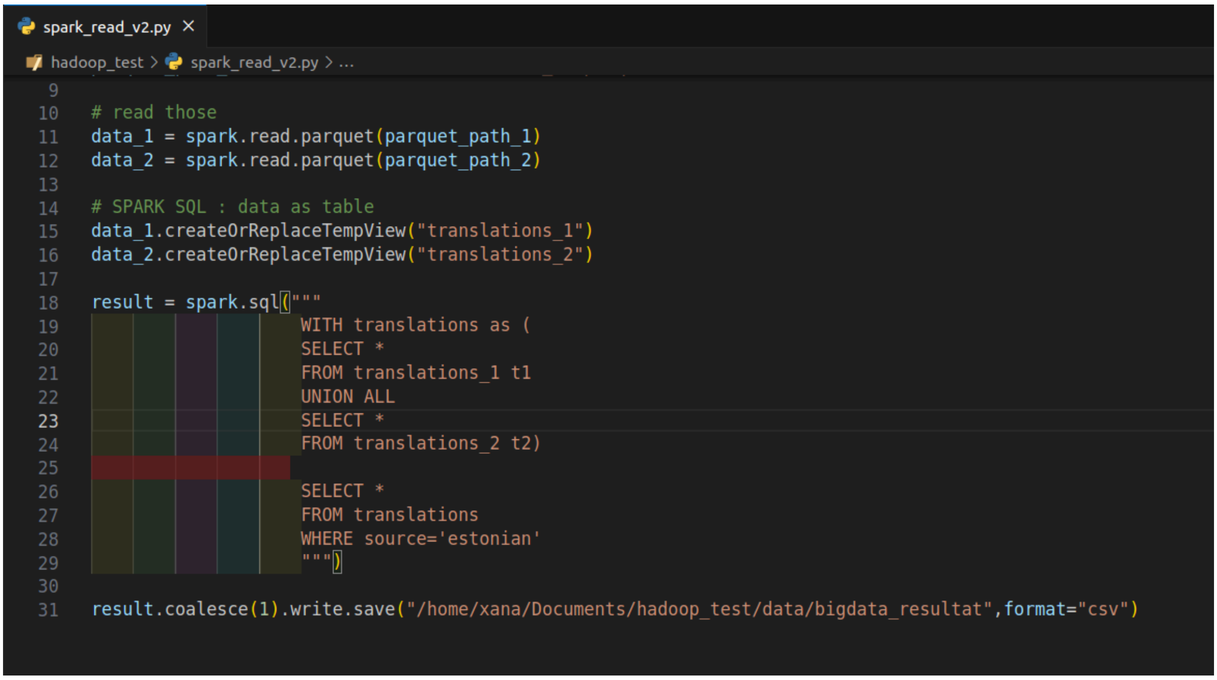
On vérifie ensuite sa présence :



```
(spark) xana@xana:~/Documents/hadoop_test$ hadoop fs -ls /
Found 2 items
-rw-r--r--   3 xana supergroup     733970 2023-12-23 15:20 /bigdata.parquet
drwxr-xr-x   - xana supergroup          0 2023-12-23 15:39 /data
```

1. Application Spark

1. Ecriture de l’application

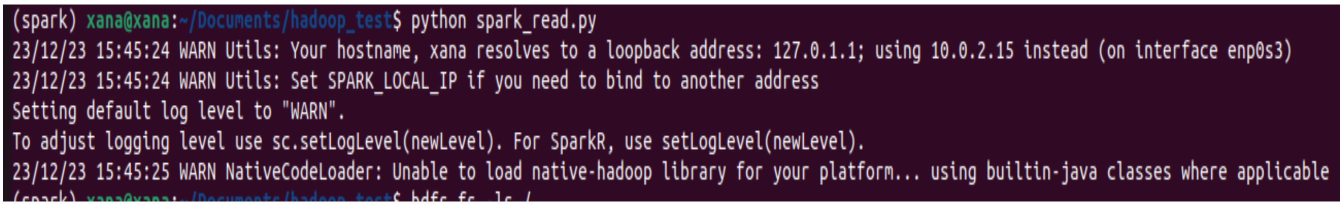


```
# read those
data_1 = spark.read.parquet(parquet_path_1)
data_2 = spark.read.parquet(parquet_path_2)

# SPARK SQL : data as table
data_1.createOrReplaceTempView("translations_1")
data_2.createOrReplaceTempView("translations_2")

result = spark.sql("""
                    WITH translations as (
                    SELECT *
                    FROM translations_1 t1
                    UNION ALL
                    SELECT *
                    FROM translations_2 t2)

                    SELECT *
                    FROM translations
                    WHERE source='estonian'
                    """)

result.coalesce(1).write.save("/home/xana/Documents/hadoop_test/data/bigdata_resultat",format="csv")
```

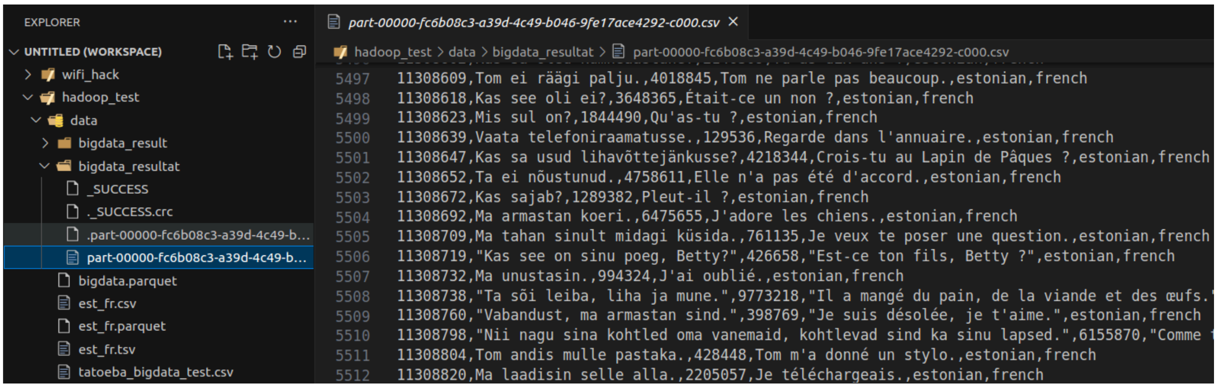
Pour complexifier un tant soit peu la requête SQL , on a ajouté d’autres données.

On sauvegarde le résultat dans le dossier désiré au format csv

2. Exécution et vérification du résultat



```
(spark) xana@xana:~/Documents/hadoop_test$ python spark_read.py
23/12/23 15:45:24 WARN Utils: Your hostname, xana resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
23/12/23 15:45:24 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/12/23 15:45:25 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
(spark) xana@xana:~/Documents/hadoop_test$ hdfs fs -ls /
```


EXPLORER
UNTITLED (WORKSPACE)
wifi_hack
hadoop_test
data
bigdata_result
bigdata_resultat
_SUCCESS
._SUCCESS.crc
.part-00000-fc6b08c3-a39d-4c49-b...
part-00000-fc6b08c3-a39d-4c49-b...
bigdata.parquet
est_fr.csv
est_fr.parquet
est_fr.tsv
tatoeba_bigdata_test.csv
part-00000-fc6b08c3-a39d-4c49-b046-9fe17ace4292-c000.csv
hadoop_test > data > bigdata_resultat > part-00000-fc6b08c3-a39d-4c49-b046-9fe17ace4292-c000.csv

```
11308609,Tom ei räägi palju.,4018845,Tom ne parle pas beaucoup.,estonian,french
11308618,Kas see oli ei?,3648365,Était-ce un non ?,estonian,french
11308623,Mis sul on?,1844490,Qu'as-tu ?,estonian,french
11308639,Vaata telefoniraamatusse.,129536,Regarde dans l'annuaire.,estonian,french
11308647,Kas sa usud lihavõttejänkusse?,4218344,Crois-tu au Lapin de Pâques ?,estonian,french
11308652,Ta ei nõustunud.,4758611,Elle n'a pas été d'accord.,estonian,french
11308672,Kas sajab?,1289382,Pleut-il ?,estonian,french
11308692,Ma armastan koeri.,6475655,J'adore les chiens.,estonian,french
11308709,Ma tahan sinult midagi küsida.,761135,Je veux te poser une question.,estonian,french
11308719,"Kas see on sinu poeg, Betty?",426658,"Est-ce ton fils, Betty ?",estonian,french
11308732,Ma unustasin.,994324,J'ai oublié.,estonian,french
11308738,"Ta sõi leiba, liha ja mune.",9773218,"Il a mangé du pain, de la viande et des œufs."
11308760,"Vabandust, ma armastan sind.",398769,"Je suis désolée, je t'aime.",estonian,french
11308798,"Nii nagu sina kohtled oma vanemaid, kohtlevad sind ka sinu lapsed.",6155870,"Comme t
11308804,Tom andis mulle pastaka.,428448,Tom m'a donné un stylo.,estonian,french
11308820,Ma laadisin selle alla.,2205057,Je téléchargeais.,estonian,french
```