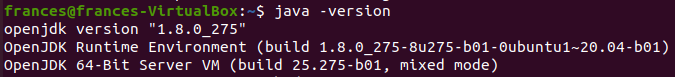
How to implement MapReduce using Hadoop file system?

Step 1: Environment

First you should make sure Java and Hadoop are properly installed on your Linux or Virtual Machine.

Use this command to check Java version.

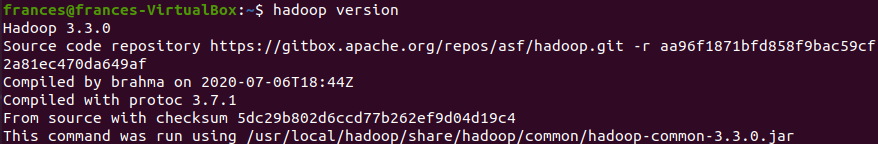
$ java -version



Make sure your Java is 1.8 version. Other versions may not support.

Use this command to check Hadoop version.

$ hadoop version



My version is 3.3.0. Please make sure to use the same version as mine.

Pull the GitHub repository using this command:

$ git clone https://github.com/Annalina-Luo/Text-Search-Engine-based-on-Hadoop

Move this file to your home directory and go to this file using:

$ cd Text-Search-Engine-based-on-Hadoop/code

Step 2: Compile Java code

Use this command to compile Java code I provide:

$ hadoop com.sun.tools.javac.Main TFIDF\_InvertedIndex.java

You may encounter an error:

Error: Could not find or load main class com.sun.tools.javac.Main

Here is the command to fix it:

$ sudo vim ~/.bashrc

Enter your password and press “Enter”.

In the bashrc file, you should check if your environmental variables are set correctly, especially for this line:

export HADOOP\_CLASSPATH=$JAVA\_HOME/lib/tools.jar

If you modified it, restart the terminal, or use following command to apply the changes:

$ source ~/.bashrc

Run the compile command again and you should get four .class files.

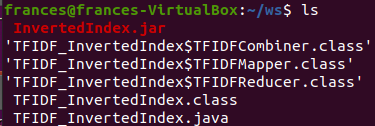
Then create a jar file using the following command:

$ jar cf InvertedIndex.jar TFIDF\_InvertedIndex\*.class

Check if it is created successfully.

$ ls

It should look like this:



Step 3: Use Hadoop to MapReduce

Use the following command to start Hadoop system:

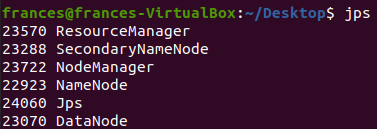
$ start-dfs.sh

$ start-yarn.sh

Use this command to check whether it start normally:

$ jps

It should look like this:



Create an input directory in Hadoop, for example a directory called input:

$ hdfs dfs -mkdir -p /usr/local/hadoop/hdfs/input/

You might get an error like this:

Error: Your namenode is in safe mode.

You can type the following command to exit safe mode:

$ hdfs dfsadmin -safemode leave

Create a subdirectory under input, for example a subdirectory called invertedIndex:

$ hdfs dfs -mkdir -p /usr/local/hadoop/hdfs/input/invertedIndex/

Then go to file directory and unzip the text file we collected.

$ cd ~/Text-Search-Engine-based-on-Hadoop/file

$ tar –xzvf files.zip

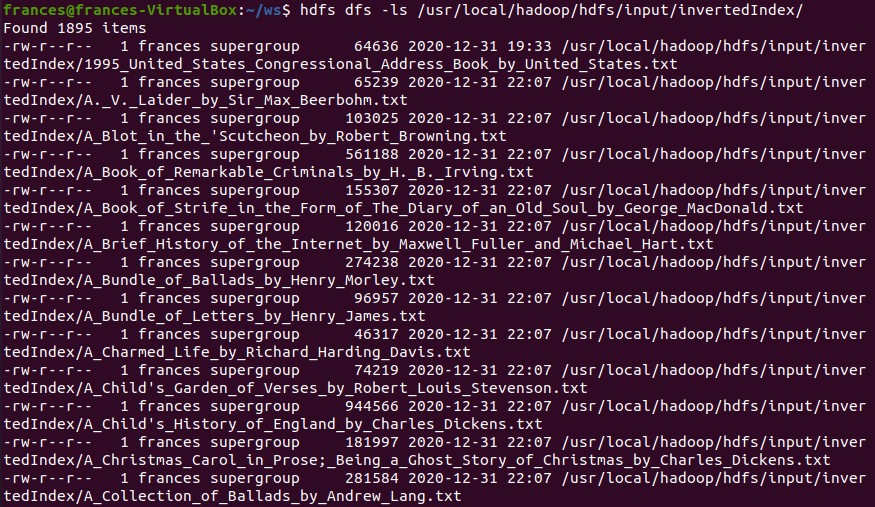
Then upload the text documents from the directory

$ hdfs dfs -put ~/Text-Search-Engine-based-on-Hadoop/file/files/\* /usr/local/hadoop/hdfs/input/invertedIndex/

Check if the upload is successful:

$ hdfs dfs -ls /usr/local/hadoop/hdfs/input/invertedIndex/

It should look like this:

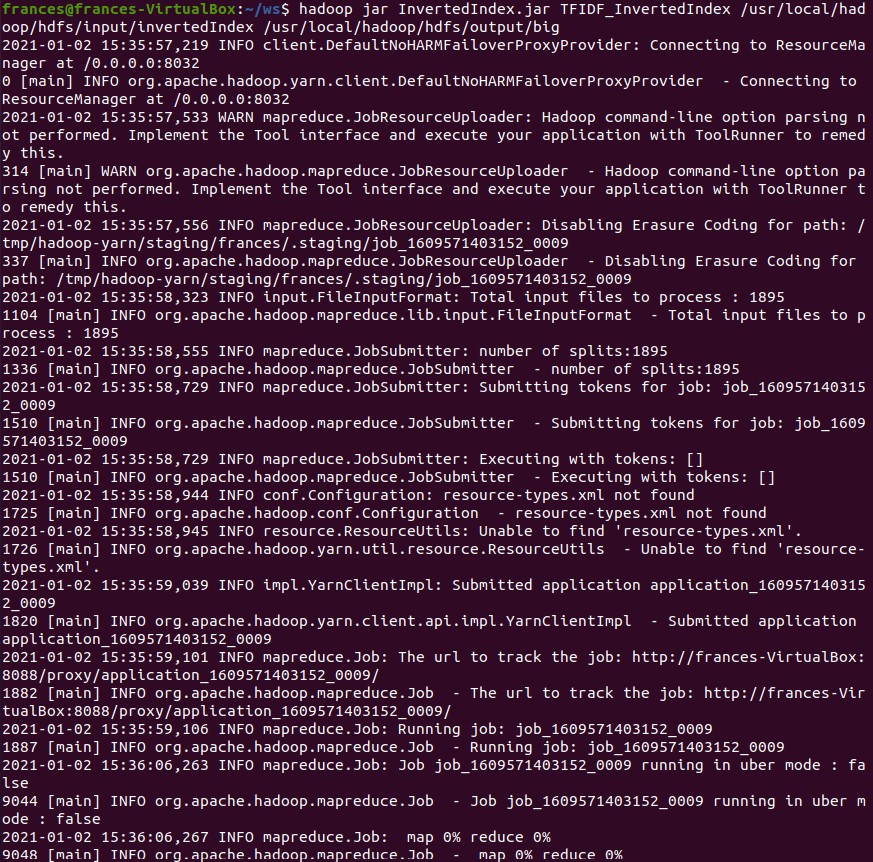


Run the application using the command:

$ hadoop jar InvertedIndex.jar TFIDF\_InvertedIndex /usr/local/hadoop/hdfs/input/invertedIndex /usr/local/hadoop/hdfs/output/big

You can change output directory to any name you like, for me I use “big”.

It should start to run MapReduce job:

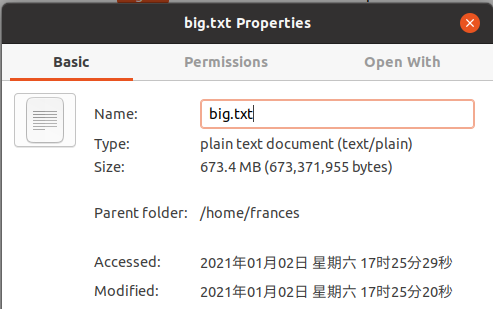


Use this command to copy result to your local file system:

$ hdfs dfs -get /usr/local/hadoop/hdfs/output/big/part-r-00000 ~/big.txt

You can change the text filename as you like.

Then you have the .txt file in the home directory.



Don’t forget to stop Hadoop using:

$ stop-dfs.sh

$ stop-yarn.sh