Hadoop 3.x installation guide

Version 3 and later of Apache Hadoop requires at least Java 8. Here we will use Hadoop 3.2.0 and Java 8. You are free to use other versions.

1. Install Java:

    $sudo apt-get update
    $sudo apt-get install openjdk-8-jre openjdk-8-jdk

    Set Java_Home:

    $vi ~/.bashrc
    Add the following command to the file:

    export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
    Save and close the file. Activate the setting immediately with:

    $source ~/.bashrc

2. Install Hadoop
    Download Hadoop-3.2.0.tar.gz:

    $sudo wget http://mirrors.advancedhosters.com/apache/hadoop/common/hadoop-3.2.0/hadoop-3.2.0.tar.gz

    Install Hadoop to /usr/local
    $ sudo tar -zxvf ./hadoop-3.2.0.tar.gz -C /usr/local
    $ cd /usr/local/
    $ sudo mv ./hadoop-3.2.0 ./hadoop
    $ sudo chown -R username:username ./hadoop   (This command is to change permission of hadoop file. Replace the username with your account username. For example, my account username is Fengli and the command will be like: $ sudo chown -R fengli:fengli ./hadoop)

3. Run an example of Hadoop
    $cd /usr/local/hadoop
    $mkdir input
    $cp ./etc/hadoop/*.xml input
    $./bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar grep input output 'dfs[a-z.]+'
    $cat ./output/part-r-00000
    $rm -R ./output   (delete output)
    Result:

```
            Shuffled Maps =1
            Failed Shuffles=0
            Merged Map outputs=1
            GC time elapsed (ms)=66
            Total committed heap usage (bytes)=307249152
      Shuffle Errors
            BAD_ID=0
            CONNECTION=0
            IO_ERROR=0
            WRONG_LENGTH=0
            WRONG_MAP=0
            WRONG_REDUCE=0
      File Input Format Counters
            Bytes Read=123
      File Output Format Counters
            Bytes Written=23
fengli@fengli-VirtualBox:/usr/local/hadoop$ cat ./output/part-r-00000
1       dfsadmin
fengli@fengli-VirtualBox:/usr/local/hadoop$
```

Run MapReduce "WordCount" program

1. Add CLASSPATH to bashrc file.

    $ vi   ~/.bashrc

  Set the Hadoop path by adding the following commands to the file:

    HADOOP_HOME=/usr/local/hadoop
    export CLASSPATH=$HADOOP_HOME/share/hadoop/common/hadoop-common-
    3.2.0.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-core-
    3.2.0.jar:$HADOOP_HOME/share/hadoop/common/lib/commons-cli-1.2.jar:$CLASSPATH

  Activate the path setting immediately:

    $source ~/.bashrc

2. Compile WordCount program(ignore the Warnings)

    $javac WordCount.java -Xlint:deprecation

3. Make a jar file

    $jar -cvf WordCount.jar ./WordCount*.class

4. Set up input files

    $mkdir input
    $echo "echo of the rainbow" > ./input/file0
    $echo "the waiting game" > ./input/file1

5. Run the program

    $/usr/local/hadoop/bin/hadoop jar WordCount.jar WordCount input output

Result:

```
                Reduce input groups=6
                Reduce shuffle bytes=91
                Reduce input records=7
                Reduce output records=6
                Spilled Records=14
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=89
                Total committed heap usage (bytes)=555233280
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=37
        File Output Format Counters
                Bytes Written=57
fengli@fengli-VirtualBox:~/cloudComputing/p1$ cat ./output/part-r-00000
echo    1
game    1
of      1
rainbow 1
the     2
waiting 1
fengli@fengli-VirtualBox:~/cloudComputing/p1$
```