Group Members: Malay Agarwal ,

Umair Ahmad Beig,

Sami Ullah Naikoo.

Question 2:

Implement a pair of a Map and a Reduce function which, for each distinct term that occurs in any of the text documents in Wikipedia-EN-20120601 ARTICLES.tar.gz, counts the number of distinct documents in which the term appears. We will call this value the Document Frequency (DF) of that term in the entire set of Wikipedia articles. Store the resulting DF values of all terms in a single TSV file with the following schema: TERMDF While generating the output in the above format, consider filtering out all terms that belong to the stopwords.txt file shared on LMS. (You may perform this filter operation in your map method.) Identify the top 100 terms with a high document frequency. Use those terms alone for the next sub problem.

Steps:

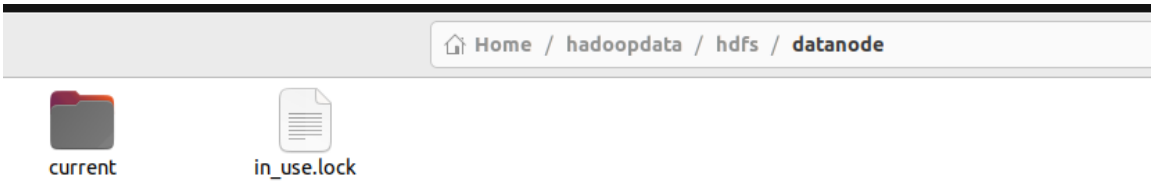
1. Create a maven project.
2.
3. Add the required Dependencies in pom.xml
4. Hadoop core

5. Add maven assembly plugin for making fat jar ie JAR-WITH-DEPENDENCIES in pom.xml
6.
7. Add the DocumentFrequency class in the com.example package.

8. From the terminal “mvn clean install” to generate the required jar with dependencies file.

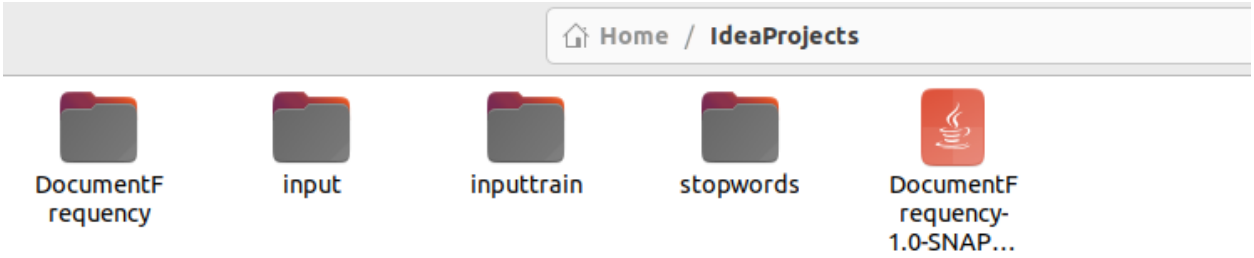
9. Move the jar-with-dep from target to directory easily accessible(IdeaProjects dir in my case).

10. Clearing the hadoop datanode directory
    a. Delete the files in datanode dir
        i.
        ii.



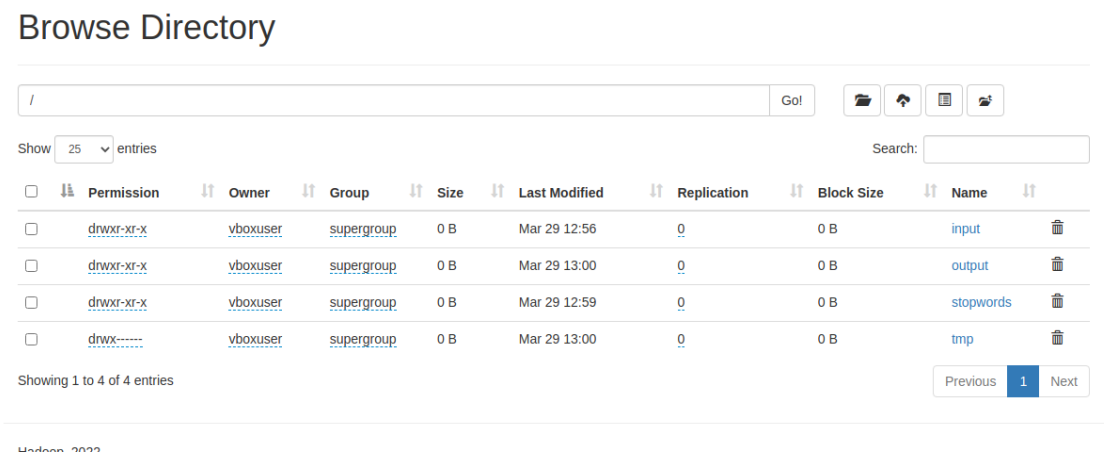
    b. Stop-all.sh
    c. hdfs namenode -format
    d. Start-all.sh
    e. Jps (to check the status)

11. In my case the local dirs alongside my project dir looked like this:



12. Now we need to make an input and stopword directories in hadoop fs.
    - hdfs dfs -mkdir /input
    - hdfs dfs -mkdir /stopwords

13. Push the documents in inputtrain folder into the /input and stopwords/stopwords.txt into /stopwords using “hdfs dfs -put” command.



14. Now we need to run the hadoop jar <jarname> /input /output /stopwords/stopwords.txt

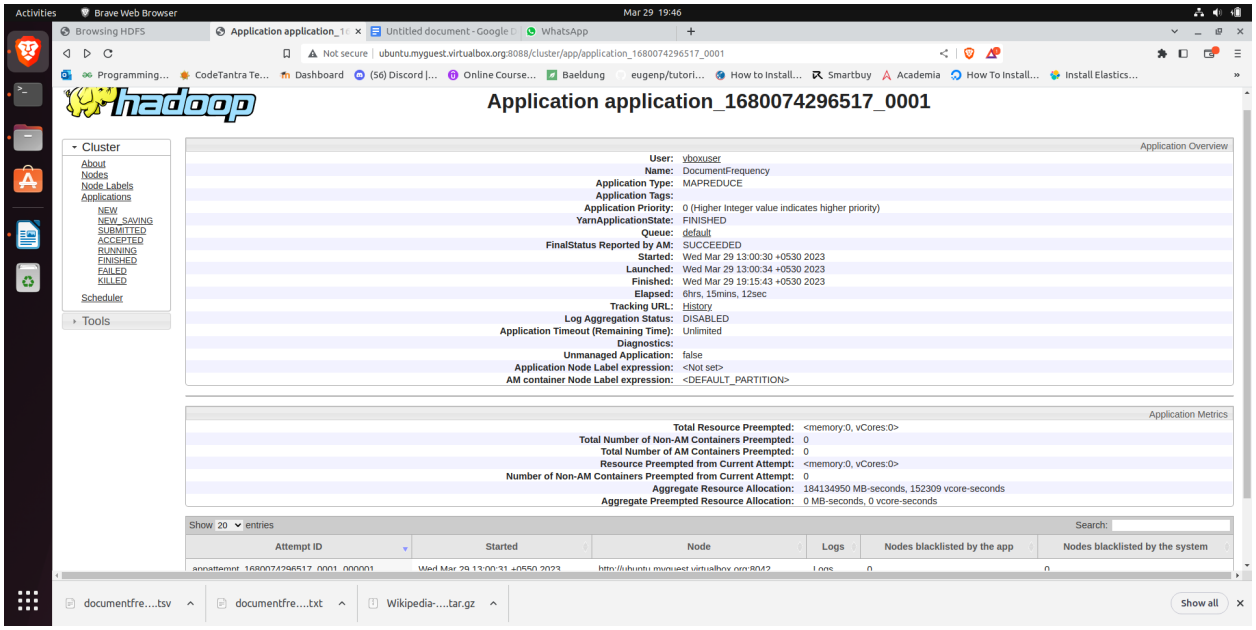
To sort the output file and take top 100 pairs:
From CLI:

hdfs dfs -cat /output/part-r-00000|sort -k 2 -n -r|head -n 100|hdfs dfs -put - /output/documentfreqoutputFULL.tsv

k===column 2
n === numeric
r===reverse

****
****

Screenshots related to the progress of mapreduce :


Activities
Brave Web Browser
Mar 29 19:46
Browsing HDFS
Application application_1
Untitled document - Google D
WhatsApp
Not secure
ubuntu.myguest.virtualbox.org:8088/cluster/app/application_1680074296517_0001
Programming...
CodeTantra Te...
Dashboard
(56) Discord |...
Online Course...
Baeldung
eugenp/tutori...
How to Install...
Smartbuy
Academia
How To Install...
Install Elastics...
hadoop
Application application_1680074296517_0001
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Application Overview
User: vboxuser
Name: DocumentFrequency
Application Type: MAPREDUCE
Application Tags:
Application Priority: 0 (Higher Integer value indicates higher priority)
YarnApplicationState: FINISHED
Queue: default
FinalStatus Reported by AM: SUCCEEDED
Started: Wed Mar 29 13:00:30 +0530 2023
Launched: Wed Mar 29 13:00:34 +0530 2023
Finished: Wed Mar 29 19:15:43 +0530 2023
Elapsed: 6hrs, 15mins, 12sec
Tracking URL: History
Log Aggregation Status: DISABLED
Application Timeout (Remaining Time): Unlimited
Diagnostics:
Unmanaged Application: false
Application Node Label expression: <Not set>
AM container Node Label expression: <DEFAULT_PARTITION>
Application Metrics
Total Resource Preempted: <memory:0, vCores:0>
Total Number of Non-AM Containers Preempted: 0
Total Number of AM Containers Preempted: 0
Resource Preempted from Current Attempt: <memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt: 0
Aggregate Resource Allocation: 184134950 MB-seconds, 152309 vcore-seconds
Aggregate Preempted Resource Allocation: 0 MB-seconds, 0 vcore-seconds
Show 20 entries
Search:
Attempt ID
Started
Node
Logs
Nodes blacklisted by the app
Nodes blacklisted by the system
appattempt_1680074296517_0001_000001
Wed Mar 29 13:00:31 +0530 2023
http://ubuntu.myguest.virtualbox.org:8042
Logs
0
0
documentfre....tsv
documentfre....txt
Wikipedia-....tar.gz
Show all