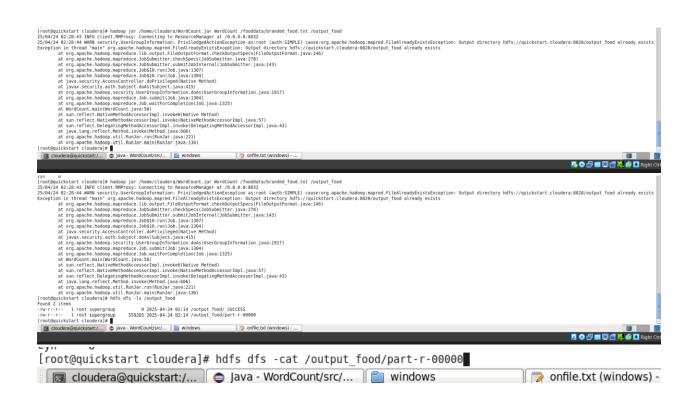
PROBLEM1:-

- IMPLEMENT A WORD COUNT PROGRAM THAT NOT ONLY COUNTS WORDS BUT ALSO FILTERS OUT STOP WORDS (LIKE "THE," "AND," "A") AND OUTPUTS THE TOP N MOST FREQUENT WORDS.
- OBJECTIVE: PRACTICE WITH BASIC TEXT PROCESSING, FILTERING, AND SORTING

USING PARALLEL AND DISTRIBUTED COMPUTING CONCEPT IN HADOOP, WRITE A PROGRAM TO COUNT HOW MANY TIMES EACH WORD IN THE DATASETS OCCURS AND SHOW HOW PARALLEL COMPUTING CONCEPTS IS APPLIED HERE. HOW MANY BLOCKS IS CREATED IN THIS DATASET EXPLAIN BLOCK CONCEPTS AND SHOW WITH SNAP THE NUMBER BLOCKS CREATED, SUBMIT THE LAB REPORT WITH PROPER PROOF OF EXECUTION (SNAP).

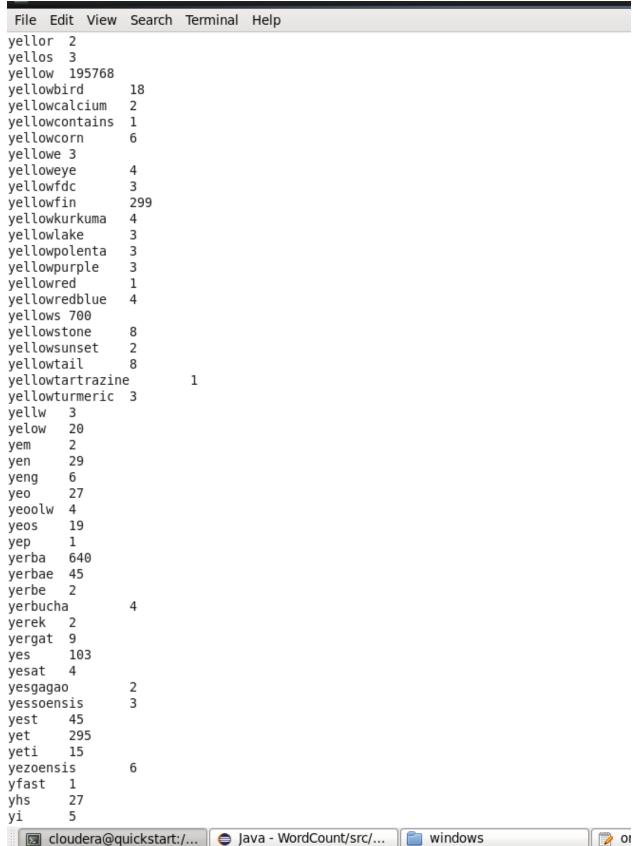
Solution:

The Code:



The Word Count Results:

The Word Count F			
File Edit View	Search	Terminal	Help
ycerides ycol 3	6		
ydrolyzed	2		
ye 9			
yea 3			
yeager 27			
yeah 27			
year 20			
years 4			
yearst 3			
yeas 26 veasi 2			
yeasi 2 yeast 178999			
yeastbacteria	3		
yeastbrewers	4		
yeastcaramel	3		
yeastchocolaty	2		
yeastcontains	7		
yeastcorn	3		
yeastdextrose	2		
yeastdry	1		
yeaste 1			
yeastexetctract			
yeastgarlic	2		
yeastj 2 yeastl 3			
yeastm 3			
yeastmolasses	1		
yeastmustard	1		
yeastnon	3		
yeastorganic	1		
yeastpacket	6		
yeasts 41			
yeastsalt	4		
yeastsobitian	1		
yeastsorbitian	13	_	
yeastspicesonio		3	
yeastvital	4		
yeastwheat	3		
yeastyeast yeat 1	3		
yeats 24			
yee 8			
yeehaw 12			
yeh 2			
yehuda 43			
yein 35			
yelin 25			
yell 4			
:: [··· ·- ·· · · · · · · · · · · · · · · ·





Top 10:

```
[root@quickstart cloudera]# hdfs dfs -cat /output food/part-r-00000 | sort -k2 -nr | head -n 10
salt
       1170825
united 1051935
states 1049305
        1000822
li
        911115
oil
        842915
sugar
       840751
acid
       830045
       707140
water
flour
       659470
[root@quickstart cloudera]#
```

Block Info for branded_food.txt:

```
[root@quickstart cloudera]# hdfs fsck /fooddata/branded_food.txt -files -blocks -locations
Connecting to namenode via http://quickstart.cloudera:50070/fsck?ugi=root&files=l&blocks=l&locations=l&path=%2Ffooddata%2Fbranded_food.txt
FSCK started by root (auth:SIMPLE) from /10.0.2.15 for path /fooddata/branded_food.txt at Thu Apr 24 02:38:34 PDT 2025
ffooddata/branded_food.txt 423280148 bytes, 4 block(s): 0K
0. BP-1067413441-127.0.0.1-15009775264580:blk 10973742786 1966 len=134217728 live repl=1 [DatanodeInfowlithStorage[10.0.2.15:50010,b5-621c9278-caa3-4a7b-bf10-3c8a1245cb51,DISK]]
2. BP-1067413441-127.0.0.1-15009775264580:blk 10973742787 1967 len=134217728 live repl=1 [DatanodeInfowlithStorage[10.0.2.15:50010,b5-621c9278-caa3-4a7b-bf10-3c8a1245cb51,DISK]]
3. BP-1067413441-127.0.0.1-15009775264580:blk 10973742789 1969 len=20626964 Live_repl=1 [DatanodeInfowlithStorage[10.0.2.15:50010,b5-621c9278-caa3-4a7b-bf10-3c8a1245cb51,DISK]]

Status: HEALTHY
Total size: 423280148 B
Total dirs: 0
Total symlinks: 0
Total spulinks: 0
Total blocks (validated): 4 (avg. block size 105820037 B)
Minimally replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks:
```