

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:

```
[root@quickstart cloudera]# hadoop jar /home/cloudera/WordCount.jar WordCount /fooddata/branded_food.txt /output_food
25/04/24 02:28:43 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
25/04/24 02:28:44 WARN security.UserGroupInformation: PrivilegedActionException as:root (auth:SIMPLE) cause:org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory hdfs://quickstart.cloudera:8020/output_food already exists
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory hdfs://quickstart.cloudera:8020/output_food already exists
    at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:146)
    at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:270)
    at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:143)
    at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1307)
    at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1304)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:415)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1917)
    at org.apache.hadoop.mapreduce.Job.submit(Job.java:1304)
    at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1325)
    at WordCount.main(WordCount.java:58)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:606)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
[root@quickstart cloudera]#
cloudera@quickstart:~$ java - WordCount/src/... windows onfile.txt (windows) - ...
zyn
[root@quickstart cloudera]# hadoop jar /home/cloudera/WordCount.jar WordCount /fooddata/branded_food.txt /output_food
25/04/24 02:28:43 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
25/04/24 02:28:44 WARN security.UserGroupInformation: PrivilegedActionException as:root (auth:SIMPLE) cause:org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory hdfs://quickstart.cloudera:8020/output_food already exists
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory hdfs://quickstart.cloudera:8020/output_food already exists
    at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:146)
    at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:270)
    at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:143)
    at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1307)
    at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1304)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:415)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1917)
    at org.apache.hadoop.mapreduce.Job.submit(Job.java:1304)
    at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1325)
    at WordCount.main(WordCount.java:58)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:606)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
[root@quickstart cloudera]# hdfs dfs -ls /output_food
Found 2 items
-rw-r--r-- 1 root supergroup 0 2025-04-24 02:14 /output_food/_SUCCESS
-rw-r--r-- 1 root supergroup 559265 2025-04-24 02:14 /output_food/part-r-000000
[root@quickstart cloudera]#
cloudera@quickstart:~$ java - WordCount/src/... windows onfile.txt (windows) - ...
[root@quickstart cloudera]# hdfs dfs -cat /output_food/part-r-000000
...
```

The Word Count Results:

File	Edit	View	Search	Terminal	Help
ycerides			6		
ycol	3				
ydrolyzed			2		
ye	9				
yea	3				
yeager	27				
yeah	27				
year	20				
years	4				
yearst	3				
yeas	26				
yeasi	2				
yeast	178999				
yeastbacteria	3				
yeastbrewers	4				
yeastcaramel	3				
yeastchocolaty	2				
yeastcontains	7				
yeastcorn	3				
yeastdextrose	2				
yeastdry	1				
yeaste	1				
yeastexetctract	2				
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				
yeastspicesonion			3		
yeastvital	4				
yeastwheat	3				
yeastyeast	3				
yeat	1				
yeats	24				
yee	8				
yeehaw	12				
yeh	2				
yehuda	43				
yein	35				
yelin	25				
yell	4				

```
File Edit View Search Terminal Help
yellor 2
yellos 3
yellow 195768
yellowbird 18
yellowcalcium 2
yellowcontains 1
yellowcorn 6
yellowe 3
yelloweye 4
yellowfdc 3
yellowfin 299
yellowkurkuma 4
yellowlake 3
yellowpolenta 3
yellowpurple 3
yellowred 1
yellowredblue 4
yellows 700
yellowstone 8
yellowsunset 2
yellowtail 8
yellowtartrazine 1
yellowturmeric 3
yellw 3
yelow 20
yem 2
yen 29
yeng 6
yeo 27
yeoolw 4
yeos 19
yep 1
yerba 640
yerbae 45
yerbe 2
yerbucha 4
yerek 2
yergat 9
yes 103
yesat 4
yesgagao 2
yessoensis 3
yest 45
yet 295
yeti 15
yezoensis 6
yfast 1
yhs 27
yi 5
```

cloudera@quickstart:/... Java - WordCount/src/... windows

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VirtualBox

File Machine View Input Devices Help

Applications Places System

File Edit View Search Terminal Help

```
yogini 2
yogjrt 3
yogo 11
yogourmet 1
yogourti 12
yogovera 2
yogu 2
yoguri 7
yogurico 5
yogurll 3
yogurt 71001
yogurtchips 2
yogurtcovered 8
yogurtcultured 9
yogurtcultures 3
yogurtflavored 10
yogurtgrade 7
yogurtheat 10
yogurtii 25
yogurtpasteurized 3
yogurtpowder 3
yogurts 3
yogurtskimmed 1
yogurtstrawberry 6
yogurty 21
yogurtyogurt 1048
yogusto 1
yohari 1
yohimbine 4
yokes 1
yokey 2
yokids 3
yokois 1
yolele 4
yolita 16
yolk 11658
yolkacerola 2
yolks 18993
yolkscontains 1
yolksdried 3
yolksm 3
yolksorganic 1
yolkssalt 3
yolkssweetened 3
yolkswater 3
yellow 2
yolo 11
yols 4
yom 1
```

cloudera@quickstart:/... Java - WordCount/src/... windows onfile.txt (windows) - ...

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
li         1000822
g          911115
oil        842915
sugar     840751
acid       830045
water     707140
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=1&blocks=1&locations=1&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
/fooddata/branded_food.txt 423280148 bytes, 4 block(s): OK
0. BP-1067413441-127.0.0.1-1508775264580:blk_1073742786_1966 len=134217728 Live repl=1 [DatanodeInfoWithStorage[10.0.2.15:50010,DS-621c9e78-caa3-4a7b-bf10-3c8a1245cb51,DISK]]
1. BP-1067413441-127.0.0.1-1508775264580:blk_1073742787_1967 len=134217728 Live repl=1 [DatanodeInfoWithStorage[10.0.2.15:50010,DS-621c9e78-caa3-4a7b-bf10-3c8a1245cb51,DISK]]
2. BP-1067413441-127.0.0.1-1508775264580:blk_1073742788_1968 len=134217728 Live repl=1 [DatanodeInfoWithStorage[10.0.2.15:50010,DS-621c9e78-caa3-4a7b-bf10-3c8a1245cb51,DISK]]
3. BP-1067413441-127.0.0.1-1508775264580:blk_1073742789_1969 len=20626964 Live repl=1 [DatanodeInfoWithStorage[10.0.2.15:50010,DS-621c9e78-caa3-4a7b-bf10-3c8a1245cb51,DISK]]

Status: HEALTHY
Total size: 423280148 B
Total dirs: 0
Total files: 1
Total symlinks: 0
Total blocks (validated): 4 (avg. block size 105820037 B)
Minimally replicated blocks: 4 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Thu Apr 24 02:38:34 PDT 2025 in 79 milliseconds

The filesystem under path '/fooddata/branded_food.txt' is HEALTHY
[root@quickstart cloudera]#
```
