INFSCI 2750: Cloud Computing

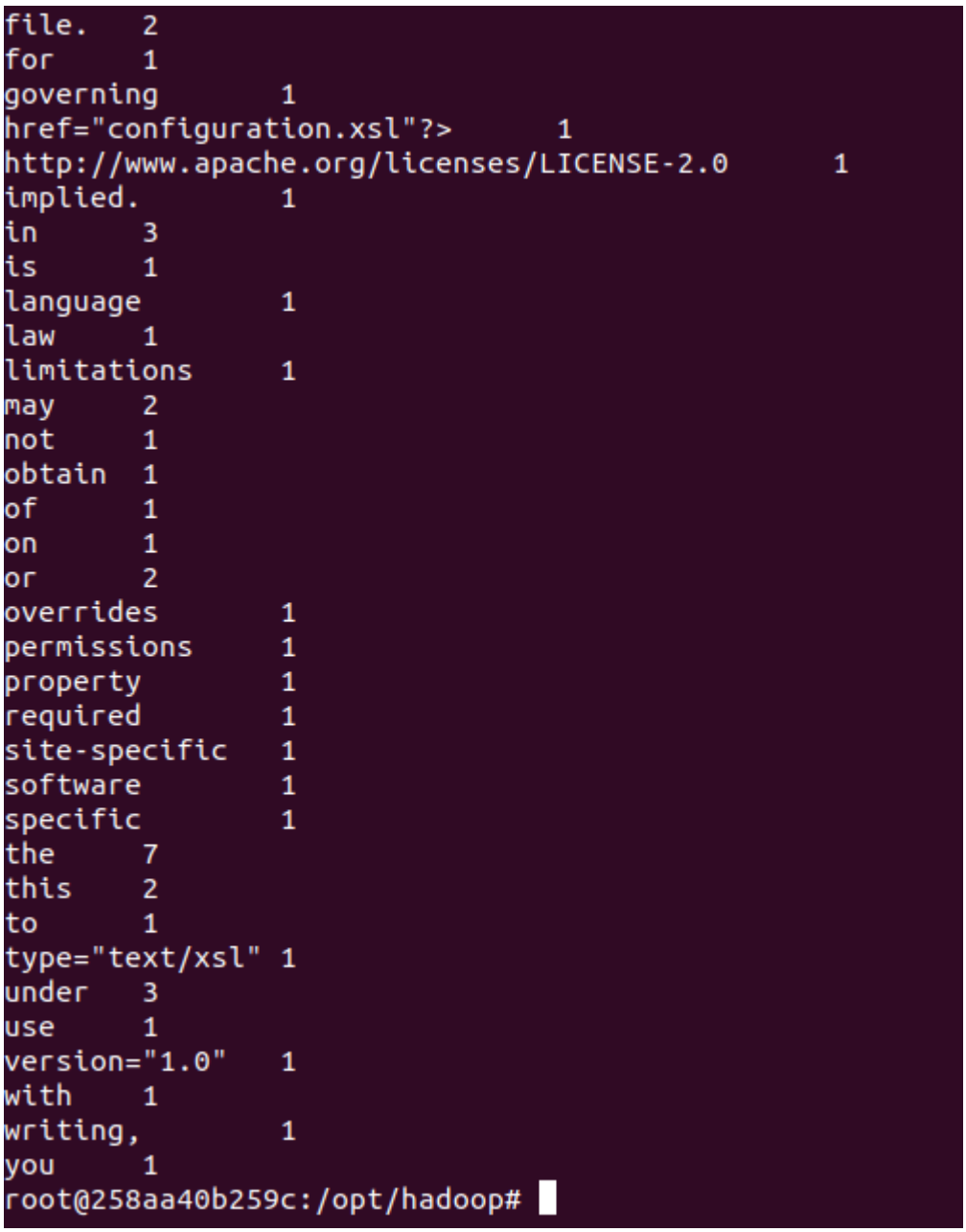
Mini Project 1

Pinhao Wang (PIW17) | Zhecheng Qi (zhq27) | Haotian Wu (haw145)

Part 1:

The wordcount example has been finished successfully, the output is:

```
root@258aa40b259c:/opt/hadoop# hdfs dfs -cat output/*
"AS     1
"License");     1
(the    1
-->     2
2.0     1
<!--    2
</configuration>        1
<?xml   1
<?xml-stylesheet        1
<configuration> 1
ANY     1
Apache  1
BASIS,  1
CONDITIONS      1
IS"     1
KIND,   1
LICENSE 1
License 3
License,        1
License.        2
Licensed        1
OF      1
OR      1
Put     1
See     2
Unless  1
Version 1
WARRANTIES      1
WITHOUT 1
You     1
a       1
accompanying    1
agreed  1
an      1
and     1
applicable      1
at      1
by      1
compliance      1
copy    1
distributed     2
either  1
encoding="UTF-8"?>      1
except  1
express 1
file    1
```



Screenshot 1. cat of the output of the worldcount example

Part 2 Hadoop program - n-gram:

A n-gram Hadoop program has been implemented to produce the n-gram frequencies of the input file with given *n* as a parameter. To test our program, we have performed n-gram with *n* equals to 2 on the input file with simple text “Helloworld”. The *n* must be passed as a parameter in args[2], otherwise an error will occur. The results are shown below:

```
ubuntu@ccprojects-22:~/hadoop$ hdfs dfs -cat output/*
2023-02-14 20:42:07,103 INFO sasl.SaslDataTransferClient: SASL encryption trust
 check: localHostTrusted = false, remoteHostTrusted = false
He      1
el      1
ld      1
ll      1
lo      1
or      1
ow      1
rl      1
wo      1
```

Screenshot 2. N-gram result with n = 2

Part 3 Hadoop program - Log analysis:

1. How many hits were made to the website item "/assets/img/home- logo.png"? Answer:

98776 hits

```
ubuntu@ccprojects-22:~/hadoop$ hdfs dfs -cat output/*
2023-02-15 00:03:49,572 INFO sasl.SaslDataTransferClient: SASL encryption trust
 check: localHostTrusted = false, remoteHostTrusted = false
/assets/img/home-logo.png       98776
```

2. How many hits were made from the IP: 10.153.239.5

Answer: 547

```
ubuntu@ccprojects-22:~/hadoop$ hdfs dfs -cat output/*
2023-02-15 00:10:45,627 INFO sasl.SaslDataTransferClient: SASL encryption trust
 check: localHostTrusted = false, remoteHostTrusted = false
10.153.239.5    547
```

3. Which path in the website has been hit most? How many hits were made to the path?

Answer: /assets/css/combined.css, with 117348 hits

```
ubuntu@ccprojects-22:~/hadoop$ hdfs dfs -cat output/*
2023-02-15 00:31:06,902 INFO sasl.SaslDataTransferClient: SASL encryption trust
 check: localHostTrusted = false, remoteHostTrusted = false
/assets/css/combined.css        117348
/assets/js/javascript_combined.js       106818
/       99299
/assets/img/home-logo.png       98744
/assets/css/printstyles.css     93158
/images/filmpics/0000/3695/Pelican_Blood_2D_Pack.jpg    91933
/favicon.ico    66831
/robots.txt     51975
/images/filmpics/0000/3139/SBX476_Vanquisher_2d.jpg     39591
/assets/img/search-button.gif   38990
```

4. Which IP accesses the website most? How many accesses were made by it?

Answer:10.216.113.172, with 158614 hits

```
ubuntu@ccprojects-22:~/hadoop$ hdfs dfs -cat output/*
2023-02-15 00:38:51,345 INFO sasl.SaslDataTransferClient: SASL encryption trust
 check: localHostTrusted = false, remoteHostTrusted = false
10.216.113.172  158614
10.220.112.1    51942
10.173.141.213  47503
10.240.144.183  43592
10.41.69.177    37554
10.169.128.121  22516
10.211.47.159   20866
10.96.173.111   19667
10.203.77.198   18878
10.31.77.18     18721
```

Note: for the questions 3 and 4, we developed Hadoop programs to get top 10 hits by IP/URL