Faculty of engineering

Assignment 1

Distributed Systems

Presented by

1-	**Seif El Din Wael Salem**	**18010832**
2-	**Mohamed Mostafa Badran**	**18011621**
3-	**Ziad Hassan Mahmoud**	**18010720**

03/4/2022

We cannot talk about distributed systems without mentioning Apache Hadoop. Hadoop has been growing vastly in the recent years and this is due to its simplicity and effectiveness. In addition, it introduces a simple map reduce framework that is used to distribute the workload on the all connected name nodes. In this report, we will discuss how to deal with Hadoop and show our output results.

Table of Contents

1 Problem Definition

It is required to be familiar with Apache Hadoop Software and learn how to setup a Pseudo-Distributed Operation. By Pseudo-Distributed we mean that all Hadoop daemons are on a single machine.

After setting up Hadoop, we need to create a program that calculates the number of occurrence of every word in one or more files using the Map Reduce algorithm.

2 Algorithms

Setting Up Hadoop

Setting up Hadoop is an easy task, after installing Hadoop and setting up the environment variables, we then need to perform the following steps:

1. Format the namenode.
2. Start the HDFS.
3. Start yarn (to be able to execute jars).

Map Reduce for WordCount

The algorithm work as follows:

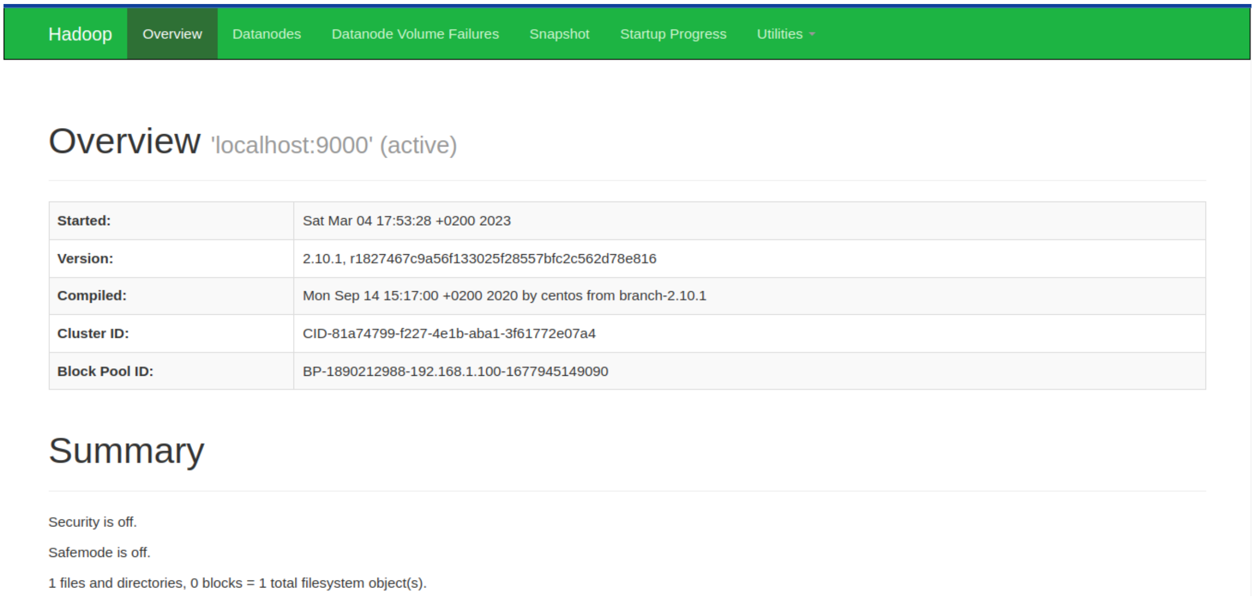
```
Create a mapper function which does the following:
    Split the record into seperate words.
    Write a value one to Context and the key is the word.
Create Reducer function:
    for each key:
        The number of occurrences of key=length of list of corresponding values
```

4 Implementation

The following implementations have been implemented on Ubuntu 22.04.

Hadoop Start up

After starting the HDFS we can show our data in the following site:



Overview 'localhost:9000' (active)

Started:	Sat Mar 04 17:53:28 +0200 2023
Version:	2.10.1, r1827467c9a56f133025f28557bfc2c562d78e816
Compiled:	Mon Sep 14 15:17:00 +0200 2020 by centos from branch-2.10.1
Cluster ID:	CID-81a74799-f227-4e1b-aba1-3f61772e07a4
Block Pool ID:	BP-1890212988-192.168.1.100-1677945149090

Summary

Security is off.

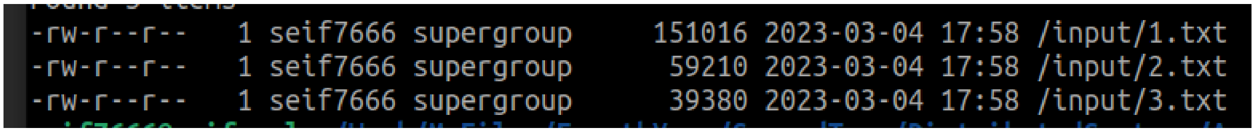
Safemode is off.

1 files and directories, 0 blocks = 1 total filesystem object(s).

Now that we are running the HDFS, we can start with the map reduce problem.

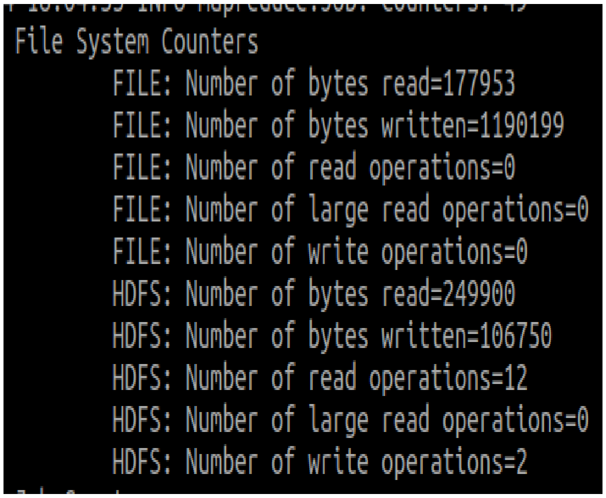
Map Reduce

After putting the input files in the HDFS:

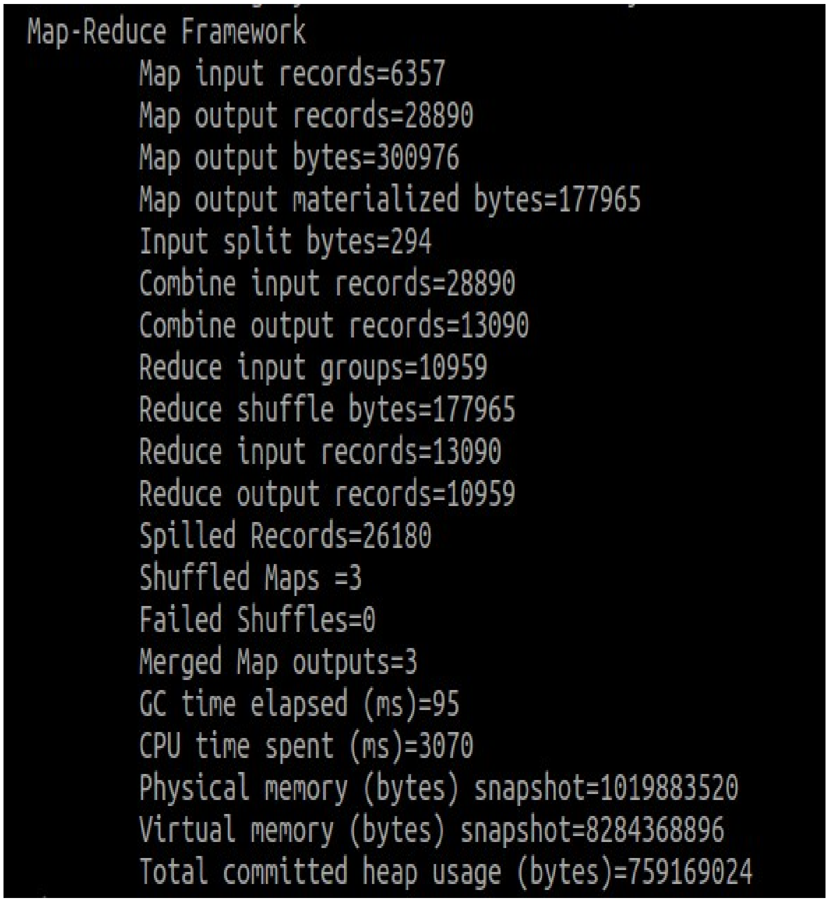


```
-rw-r--r--   1 seif7666 supergroup     151016 2023-03-04 17:58 /input/1.txt
-rw-r--r--   1 seif7666 supergroup      59210 2023-03-04 17:58 /input/2.txt
-rw-r--r--   1 seif7666 supergroup      39380 2023-03-04 17:58 /input/3.txt
```

We run the mapreduce job , we obtain the following information that shows a lot of useful information:



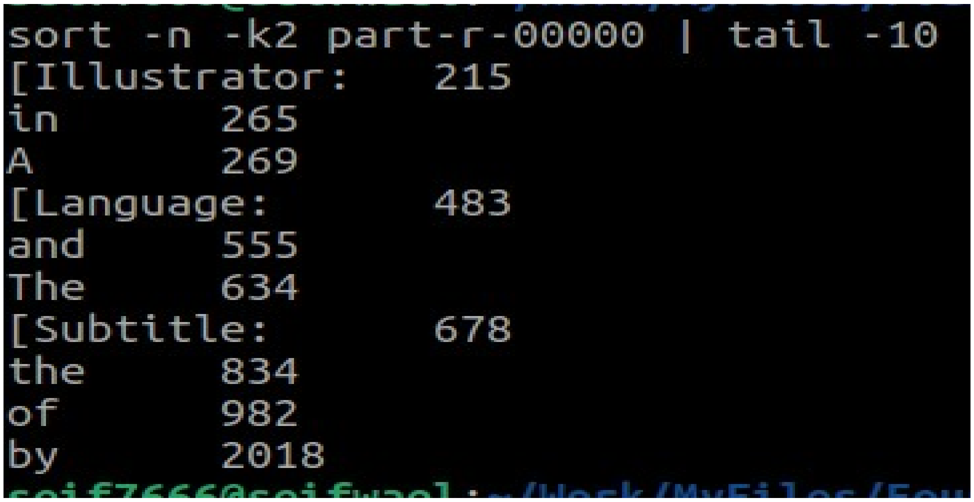
```
 File System Counters
         FILE: Number of bytes read=177953
         FILE: Number of bytes written=1190199
         FILE: Number of read operations=0
         FILE: Number of large read operations=0
         FILE: Number of write operations=0
         HDFS: Number of bytes read=249900
         HDFS: Number of bytes written=106750
         HDFS: Number of read operations=12
         HDFS: Number of large read operations=0
         HDFS: Number of write operations=2
```



```
 Map-Reduce Framework
         Map input records=6357
         Map output records=28890
         Map output bytes=300976
         Map output materialized bytes=177965
         Input split bytes=294
         Combine input records=28890
         Combine output records=13090
         Reduce input groups=10959
         Reduce shuffle bytes=177965
         Reduce input records=13090
         Reduce output records=10959
         Spilled Records=26180
         Shuffled Maps =3
         Failed Shuffles=0
         Merged Map outputs=3
         GC time elapsed (ms)=95
         CPU time spent (ms)=3070
         Physical memory (bytes) snapshot=1019883520
         Virtual memory (bytes) snapshot=8284368896
         Total committed heap usage (bytes)=759169024
```

5 Results

Now, we print the resulting output of the counted words. Since the file is too long, we will only print the first 10 lines after sorting them.



```
sort -n -k2 part-r-00000 | tail -10
[Illustrator:   215
in      265
A       269
[Language:      483
and     555
The     634
[Subtitle:      678
the     834
of      982
by      2018
seif7666@seifwael:~/Work/MyFiles/Fou
```

6 Conclusion

- Configuring Hadoop and how to check that it is running correctly.
- Implementing a simple map-reduce job and monitoring the performance.
- It’s necessary to avoid using map-reduce jobs if the we have small number of records, this is to avoid a big useless overhead computations.