CSC563 Multithreaded Distributed Programming Assignment 4: Hadoop/MapReduce

Jason Taylor
Southern Connecticut State University, Graduate Student
New Haven, Connecticut
taylorj13@southernct.edu

Abstract – I set up the Hadoop MapReduce framework to execute a word count application for this project assignment. In this assignment, I broke it down into two parts. I set up the virtual cluster of 3 nodes in the first part of the project and then installed the Hadoop framework. For the second part of the project, I ran several test applications on the Hadoop cluster that I created in step one.

Keywords—Hadoop, MapReduce

I. Introduction

In this project assignment, I used the MapReduce framework on Hadoop to run three experiments testing the Hadoop cluster. The experiments calculated Π, ran the command grep to get information from many files, and lastly, a word count application using Hadoop MapReduce. Hadoop is an open-source, Java-based framework used for storing and processing big data [2]. The data is stored on inexpensive servers that run as clusters [2]. Its distributed file system enables concurrent processing and fault tolerance [2]. MapReduce is a programming model or pattern within the Hadoop framework used to access big data stored in the Hadoop File System (HDFS) [1]. MapReduce enables concurrent processing by splitting petabytes of data into smaller chunks and processing them in parallel on Hadoop commodity servers [1]. In this assignment, I set up a virtual cluster of three nodes consisting of one master node and two worker nodes. The setup for this project was very similar to assignment 2. I just needed to set up the virtual machines, configure the networking, set up the hostnames on each node, and then connect them by updating the file /etc/hosts with each node's IP Address and hostname. One different configuration in this assignment, I needed to set up ssh keys for the hadoop user so the hadoop user could login to each of the other nodes without a password.

After the configuration steps were finished, I installed the Hadoop software and made all the necessary environment configurations to get the Hadoop software running. Some of these steps were editing many Hadoop configuration files, adding the slave nodes to the workers file, cloning the two slave nodes from the master, and changing the hostname and network settings. Once those steps were finished, I formatted the HDFS and started Hadoop. I then ran the first two experiments to verify Hadoop was running successfully. After I confirmed Hadoop was running, I moved to experiment 3. I modified the provided python file mapper.py for this experiment to remove all special characters from the word count. I added the python library for Regular expression operations [3] to this script to remove the numbers and special characters.

II. Implementation

In this section, I will discuss the implementation of this project, including a detailed setup guide that shows the configuration setting for setting up the nodes and the installation and configuration of Hadoop.

Configure Master Node

Hostname setup

```
[root@localhost ~]# echo master.localdomain > /etc/hostname
[root@localhost ~]# cat /etc/hostname
master.localdomain
[root@localhost ~]#
```

Network Setup for ifcfg-enp0s3 and ifcfg-enp0s8

Ifcfg-enp0s3:

```
[hadoop@master network-scripts]$ cat ifcfg-enp0s3
TYPE=Ethernet
PROXY_METHOD=none
BROWSER_ONLY=no
BOOTPROTO=dhcp
DEFROUTE=yes
IPV4_FAILURE_FATAL=no
IPV6INIT=yes
IPV6_AUTOCONF=yes
IPV6_DEFROUTE=yes
IPV6_FAILURE_FATAL=no
IPV6_ADDR_GEN_MODE=stable-privacy
NAME=enp0s3
UUID=772cbf25-2c82-418e-9165-7e640ab20d2d
DEVICE=enp0s3
ONBOOT=yes
[hadoop@master network-scripts]$
```

ifcfg-enp0s8

```
[hadoop@master network-scripts]$ cat ifcfg-enp0s8
TYPE=Ethernet
PROXY_METHOD=none
BROWSER_ONLY=no
BOOTPROTO=static
DEFROUTE=yes
IPV4_FAILURE_FATAL=no
IPV6INIT=yes
IPV6_AUTOCONF=yes
IPV6_DEFROUTE=yes
IPV6_FAILURE_FATAL=no
IPV6_ADDR_GEN_MODE=stable-privacy
NAME=enp0s8
IPADDR=10.0.0.120
DEVICE=enp0s8
ONBOOT=yes
[hadoop@master network-scripts]$
```

Update /ect/hosts

```
[root@localhost network-scripts]# cat /etc/hosts
127.0.0.1   localhost localhost.localdomain localhost4 localhost4.localdomain4
::1         localhost localhost.localdomain localhost6 localhost6.localdomain6

10.0.0.120 master master.localdomain
10.0.0.121 node01 node01.localdomain
10.0.0.122 node02 node02.localdomain
```

Run shutdown now -r to finalize the network settings.

Install Java on the master node.

```
[root@master ~]# yum install java-11-openjdk-devel
```

```
Installed:
  java-11-openjdk-devel.x86_64 1:11.0.13.0.8-1.el7_9
```

Verify Java

```
[root@master ~]# jps
3693 Jps
```

Install wget

```
[root@master ~]# yum install -y wget
```

```
Installed:
  wget.x86_64 0:1.14-18.el7_6.1
```

Create a Hadoop user and change the password

```
[root@master ~]# useradd hadoop
[root@master ~]# passwd hadoop
Changing password for user hadoop.
New password:
Retype new password:
passwd: all authentication tokens updated successfully.
```

Change to Hadoop user

```
[root@master ~]# su - hadoop
[hadoop@master ~]$
```

SSH configuration

```
[hadoop@master ~]$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa.
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:3y/hn25aAvFWETPMKfAUCsItg4kKdcmbCb83I+W1yco hadoop@master.localdomain
The key's randomart image is:
+---[RSA 2048]----+
| ..o.=... ..o+=o |
|. o.= +... +. ++ |
|.. o + o  o ...  |
|.   = . .  o .   |
|     + oSo. o    |
|    o = +. +.    |
|     + +  ..o..  |
|      E     o+.. |
|           .*=   |
+----[SHA256]-----+
```

Add keys as authorized key

```
[hadoop@master ~]$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
[hadoop@master ~]$ chmod 0600 ~/.ssh/authorized_keys
```

Test passwordless login

```
[hadoop@master ~]$ ssh master
Last login: Fri Dec 10 09:43:22 2021 from 10.0.0.150
```

Download and install Hadoop

```
[hadoop@master ~]$ wget http://apache.osuosl.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
--2021-12-10 09:47:20--  http://apache.osuosl.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
Resolving apache.osuosl.org (apache.osuosl.org)... 140.211.166.134, 64.50.236.52, 64.50.233.100, ...
Connecting to apache.osuosl.org (apache.osuosl.org)|140.211.166.134|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 500749234 (478M) [application/x-gzip]
Saving to: 'hadoop-3.3.0.tar.gz'

100%[======================================================================>] 500,749,234 11.5MB/s   in 3m 51s

2021-12-10 09:51:11 (2.06 MB/s) - 'hadoop-3.3.0.tar.gz' saved [500749234/500749234]

[hadoop@master ~]$ tar xfz hadoop-3.3.0.tar.gz
[hadoop@master ~]$ mv hadoop-3.3.0 hadoop
[hadoop@master ~]$ ls -l
total 489016
drwxr-xr-x. 10 hadoop hadoop       215 Jul  6  2020 hadoop
-rw-rw-r--.  1 hadoop hadoop 500749234 Jul 15  2020 hadoop-3.3.0.tar.gz
[hadoop@master ~]$
```

Setup environment variables

```
[hadoop@master ~]$ cat ~/.bashrc
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
        . /etc/bashrc
fi

# Uncomment the following line if you don't like systemctl's auto-paging feature:
# export SYSTEMD_PAGER=

# User specific aliases and functions
export HADOOP_HOME=/home/hadoop/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
```

Verify Hadoop environment variables.

```
[hadoop@master ~]$ source .bashrc
[hadoop@master ~]$ echo $HADOOP_HOME
/home/hadoop/hadoop
```

Configure Hadoop

Edit the hadoop-env.sh file to add JAVA_HOME variable

```
# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-11.0.13.0.8-1.el7_9.x86_64/
```

Test Hadoop

```
[hadoop@master hadoop]$ hadoop
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
 or    hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
  where CLASSNAME is a user-provided Java class

  OPTIONS is none or any of:
```

Edit core-site.xml

```
<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>fs.defaultFS</name>
    <value>hdfs://master:9000</value>
</property>
</configuration>
```

Edit hdfs-site.xml

```
<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/home/hadoop/hdfs/datanode</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/home/hadoop/hdfs/namenode</value>
  </property>
</configuration>
```

Edit mapred-site.xml

```
<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
     <name>mapreduce.map.memory.mb</name>
     <value>256</value>
  </property>
  <property>
     <name>mapreduce.reduce.memory.mb</name>
     <value>256</value>
  </property>
</configuration>
```

Edit yarn-site.xml

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/home/hadoop/hdfs/datanode</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/home/hadoop/hdfs/namenode</value>
  </property>
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>master</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
  <property>
    <name>yarn.nodemanager.env-whitelist</name>
    <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
  </property>
  <property>
    <name>yarn.nodemanager.resource.memory-mb</name>
    <value>1536</value>
  </property>
  <property>
    <name>yarn.scheduler.maximum-allocation-mb</name>
    <value>1536</value>
  </property>
  <property>
    <name>yarn.scheduler.minimum-allocation-mb</name>
    <value>128</value>
  </property>
  <property>
    <name>yarn.nodemanager.vmem-check-enabled</name>
    <value>false</value>
  </property>
</configuration>
```

Add Slave nodes

```
[hadoop@master hadoop]$ cat ~/hadoop/etc/hadoop/workers
node01
node02
```

Next shutdown the master node and clone 2 slave nodes from master. On the slave nodes update.

- /etc/hostname
- /etc/sysconfig/network-script/ifcfg-enp0s8

From the master node format

```
[hadoop@master hadoop]$ hdfs namenode -format
```

Stop the firewall on all nodes.

```
[root@master ~]# systemctl stop firewalld
[root@master ~]# systemctl status firewalld
● firewalld.service - firewalld - dynamic firewall daemon
   Loaded: loaded (/usr/lib/systemd/system/firewalld.service; enabled; vendor preset: enabled)
   Active: inactive (dead) since Fri 2021-12-10 10:13:03 EST; 17s ago
     Docs: man:firewalld(1)
  Process: 2701 ExecStart=/usr/sbin/firewalld --nofork --nopid $FIREWALLD_ARGS (code=exited, status=0/SUCCESS)
 Main PID: 2701 (code=exited, status=0/SUCCESS)
```

Start Hadoop

```
[hadoop@master ~]$ cat start_hadoop.sh
start-dfs.sh
start-yarn.sh
[hadoop@master ~]$ ./start_hadoop.sh
Starting namenodes on [master]
Starting datanodes
Starting secondary namenodes [master.localdomain]
2021-12-10 10:14:47,138 WARN util.NativeCodeLoader: Unable to load native-hadoop library
lasses where applicable
Starting resourcemanager
Starting nodemanagers
[hadoop@master ~]$
```

III. Test

This section of the paper will cover testing the functionality of the Hadoop cluster and displaying the results of the three experiments.

Create HDFS directory

```
[hadoop@master ~]$ hdfs dfs -mkdir -p assignment4
2021-12-10 10:31:45,483 WARN util.NativeCodeLoader: Unable to load native-hadoop library
lasses where applicable
[hadoop@master ~]$
```

Test moving files to HDFS directory

```
[hadoop@master ~]$ hdfs dfs -put alice_in_wonderland.txt assignment4
2021-12-10 10:37:13,804 WARN util.NativeCodeLoader: Unable to load nat
lasses where applicable
[hadoop@master ~]$
```

List contents of the HDFS directory

```
[hadoop@master ~]$ hdfs dfs -ls assignment4
2021-12-10 10:41:06,031 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
lasses where applicable
Found 2 items
-rw-r--r--   2 hadoop supergroup      74726 2021-12-10 10:37 assignment4/alice_in_wonderland.txt
-rw-r--r--   2 hadoop supergroup         15 2021-12-10 10:40 assignment4/hello_world.txt
[hadoop@master ~]$
```

Print content of the file from the HDFS directory

```
[hadoop@master ~]$ hdfs dfs -cat assignment4/hello_world.txt
2021-12-10 10:44:20,038 WARN util.NativeCodeLoader: Unable to
lasses where applicable
Hello World!!!
[hadoop@master ~]$
```

Remove the file from the HDFS directory

```
[hadoop@master ~]$ hdfs dfs -rm assignment4/hello_world.txt
2021-12-10 10:47:52,302 WARN util.NativeCodeLoader: Unable t
lasses where applicable
Deleted assignment4/hello_world.txt
[hadoop@master ~]$
```

Experiment 1 (Calculate PI)

```
[hadoop@master ~]$ hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar pi 30 100
Number of Maps  = 30
Samples per Map = 100
2021-12-10 10:51:56,885 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform
lasses where applicable
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3
Wrote input for Map #4
Wrote input for Map #5
Wrote input for Map #6
Wrote input for Map #7
Wrote input for Map #8
Wrote input for Map #9
Wrote input for Map #10
Wrote input for Map #11
Wrote input for Map #12
Wrote input for Map #13
Wrote input for Map #14
Wrote input for Map #15
Wrote input for Map #16
Wrote input for Map #17
Wrote input for Map #18
Wrote input for Map #19
Wrote input for Map #20
Wrote input for Map #21
Wrote input for Map #22
Wrote input for Map #23
Wrote input for Map #24
Wrote input for Map #25
Wrote input for Map #26
Wrote input for Map #27
Wrote input for Map #28
Wrote input for Map #29
Starting Job
2021-12-10 10:52:07,960 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at m
2021-12-10 10:52:08,462 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop
aging/job_1639149300891_0001
2021-12-10 10:52:09,031 INFO input.FileInputFormat: Total input files to process : 30
2021-12-10 10:52:10,016 INFO mapreduce.JobSubmitter: number of splits:30
2021-12-10 10:52:10,743 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1639149300891_0001
2021-12-10 10:52:10,743 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-12-10 10:52:11,003 INFO conf.Configuration: resource-types.xml not found
2021-12-10 10:52:11,003 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-12-10 10:52:11,457 INFO impl.YarnClientImpl: Submitted application application_1639149300891_0001
2021-12-10 10:52:11,498 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application
2021-12-10 10:52:11,499 INFO mapreduce.Job: Running job: job_1639149300891_0001
2021-12-10 10:52:21,834 INFO mapreduce.Job: Job job_1639149300891_0001 running in uber mode : false
2021-12-10 10:52:21,835 INFO mapreduce.Job:  map 0% reduce 0%
2021-12-10 10:52:49,109 INFO mapreduce.Job:  map 13% reduce 0%
2021-12-10 10:52:50,165 INFO mapreduce.Job:  map 20% reduce 0%
2021-12-10 10:53:18,374 INFO mapreduce.Job:  map 37% reduce 0%
2021-12-10 10:53:28,455 INFO mapreduce.Job:  map 37% reduce 12%
2021-12-10 10:53:35,515 INFO mapreduce.Job:  map 47% reduce 12%
2021-12-10 10:53:37,540 INFO mapreduce.Job:  map 53% reduce 12%
2021-12-10 10:53:40,562 INFO mapreduce.Job:  map 53% reduce 18%
2021-12-10 10:53:50,615 INFO mapreduce.Job:  map 60% reduce 18%
2021-12-10 10:53:52,628 INFO mapreduce.Job:  map 63% reduce 20%
2021-12-10 10:53:55,648 INFO mapreduce.Job:  map 70% reduce 20%
2021-12-10 10:53:58,670 INFO mapreduce.Job:  map 70% reduce 23%
2021-12-10 10:54:07,735 INFO mapreduce.Job:  map 77% reduce 23%
2021-12-10 10:54:10,749 INFO mapreduce.Job:  map 77% reduce 26%
2021-12-10 10:54:12,759 INFO mapreduce.Job:  map 83% reduce 26%
2021-12-10 10:54:14,782 INFO mapreduce.Job:  map 87% reduce 26%
2021-12-10 10:54:16,791 INFO mapreduce.Job:  map 87% reduce 29%
2021-12-10 10:54:23,850 INFO mapreduce.Job:  map 90% reduce 29%
2021-12-10 10:54:24,858 INFO mapreduce.Job:  map 93% reduce 29%
2021-12-10 10:54:26,879 INFO mapreduce.Job:  map 100% reduce 29%
2021-12-10 10:54:28,893 INFO mapreduce.Job:  map 100% reduce 100%
2021-12-10 10:54:30,945 INFO mapreduce.Job: Job job_1639149300891_0001 completed successfully
2021-12-10 10:54:31,058 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=666
                FILE: Number of bytes written=8191361
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=7910
                HDFS: Number of bytes written=215
                HDFS: Number of read operations=125
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=30
                Launched reduce tasks=1
                Data-local map tasks=30
                Total time spent by all maps in occupied slots (ms)=1129080
                Total time spent by all reduces in occupied slots (ms)=193060
                Total time spent by all map tasks (ms)=564540
                Total time spent by all reduce tasks (ms)=96530
                Total vcore-milliseconds taken by all map tasks=564540
                Total vcore-milliseconds taken by all reduce tasks=96530
                Total megabyte-milliseconds taken by all map tasks=144522240
                Total megabyte-milliseconds taken by all reduce tasks=24711680
        Map-Reduce Framework
                Map input records=30
                Map output records=60
                Map output bytes=540
                Map output materialized bytes=840
                Input split bytes=4370
                Combine input records=0
                Combine output records=0
                Reduce input groups=2
                Reduce shuffle bytes=840
                Reduce input records=60
                Reduce output records=0
                Spilled Records=120
                Shuffled Maps =30
                Failed Shuffles=0
                Merged Map outputs=30
                GC time elapsed (ms)=2495
                CPU time spent (ms)=10790
                Physical memory (bytes) snapshot=6971457536
                Virtual memory (bytes) snapshot=63374319616
                Total committed heap usage (bytes)=4453896192
                Peak Map Physical memory (bytes)=234713088
                Peak Map Virtual memory (bytes)=2047500288
                Peak Reduce Physical memory (bytes)=151576576
                Peak Reduce Virtual memory (bytes)=2055413760
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=3540
        File Output Format Counters
                Bytes Written=97
Job Finished in 143.255 seconds
Estimated value of Pi is 3.14133333333333333333
```

Experiment 2 (grep information from files)

```
[hadoop@master ~]$ hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar grep input output 'dfs[a-z.]+'
2021-12-10 11:03:13,511 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
lasses where applicable
2021-12-10 11:03:14,418 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/10.0.0.120:8032
2021-12-10 11:03:14,971 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.s
aging/job_1639149300891_0002
2021-12-10 11:03:15,538 INFO input.FileInputFormat: Total input files to process : 10
2021-12-10 11:03:16,000 INFO mapreduce.JobSubmitter: number of splits:10
2021-12-10 11:03:16,554 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1639149300891_0002
2021-12-10 11:03:16,554 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-12-10 11:03:16,775 INFO conf.Configuration: resource-types.xml not found
2021-12-10 11:03:16,775 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-12-10 11:03:16,822 INFO impl.YarnClientImpl: Submitted application application_1639149300891_0002
2021-12-10 11:03:16,863 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1639149300891_0002/
2021-12-10 11:03:16,863 INFO mapreduce.Job: Running job: job_1639149300891_0002
2021-12-10 11:03:24,178 INFO mapreduce.Job: Job job_1639149300891_0002 running in uber mode : false
2021-12-10 11:03:24,179 INFO mapreduce.Job:  map 0% reduce 0%
2021-12-10 11:03:47,445 INFO mapreduce.Job:  map 48% reduce 0%
2021-12-10 11:03:48,457 INFO mapreduce.Job:  map 60% reduce 0%
2021-12-10 11:04:10,632 INFO mapreduce.Job:  map 100% reduce 0%
2021-12-10 11:04:12,684 INFO mapreduce.Job:  map 100% reduce 100%
2021-12-10 11:04:14,703 INFO mapreduce.Job: Job job_1639149300891_0002 completed successfully
2021-12-10 11:04:14,788 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=205
                FILE: Number of bytes written=2907302
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=32627
                HDFS: Number of bytes written=219
                HDFS: Number of read operations=35
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=11
                Launched reduce tasks=1
                Data-local map tasks=11
                Total time spent by all maps in occupied slots (ms)=435624
                Total time spent by all reduces in occupied slots (ms)=46174
                Total time spent by all map tasks (ms)=217812
                Total time spent by all reduce tasks (ms)=23087
                Total vcore-milliseconds taken by all map tasks=217812
                Total vcore-milliseconds taken by all reduce tasks=23087
                Total megabyte-milliseconds taken by all map tasks=55759872
                Total megabyte-milliseconds taken by all reduce tasks=5910272
        Map-Reduce Framework
                Map input records=859
                Map output records=7
                Map output bytes=185
                Map output materialized bytes=259
                Input split bytes=1169
                Combine input records=7
                Combine output records=7
                Reduce input groups=4
                Reduce shuffle bytes=259
                Reduce input records=7
                Reduce output records=4
                Spilled Records=14
                Shuffled Maps =10
                Failed Shuffles=0
                Merged Map outputs=10
                GC time elapsed (ms)=1083
                CPU time spent (ms)=3820
                Physical memory (bytes) snapshot=2391883776
                Virtual memory (bytes) snapshot=22491854080
                Total committed heap usage (bytes)=1511747584
                Peak Map Physical memory (bytes)=231284736
                Peak Map Virtual memory (bytes)=2047275008
                Peak Reduce Physical memory (bytes)=124178432
                Peak Reduce Virtual memory (bytes)=2049605632
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=31458
        File Output Format Counters
                Bytes Written=219
2021-12-10 11:04:14,828 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/10.0.0.120:8032
2021-12-10 11:04:14,449 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.s
aging/job_1639149300891_0003
2021-12-10 11:04:16,762 INFO input.FileInputFormat: Total input files to process : 1
2021-12-10 11:04:17,221 INFO mapreduce.JobSubmitter: number of splits:1
2021-12-10 11:04:17,490 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1639149300891_0003
2021-12-10 11:04:17,490 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-12-10 11:04:17,534 INFO impl.YarnClientImpl: Submitted application application_1639149300891_0003
2021-12-10 11:04:17,541 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1639149300891_0003/
2021-12-10 11:04:17,541 INFO mapreduce.Job: Running job: job_1639149300891_0003
2021-12-10 11:04:29,695 INFO mapreduce.Job: Job job_1639149300891_0003 running in uber mode : false
2021-12-10 11:04:29,695 INFO mapreduce.Job:  map 0% reduce 0%
2021-12-10 11:04:35,997 INFO mapreduce.Job:  map 100% reduce 0%
2021-12-10 11:04:43,083 INFO mapreduce.Job:  map 100% reduce 100%
2021-12-10 11:04:44,103 INFO mapreduce.Job: Job job_1639149300891_0003 completed successfully
2021-12-10 11:04:44,147 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=115
                FILE: Number of bytes written=527543
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=348
                HDFS: Number of bytes written=77
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=7528
                Total time spent by all reduces in occupied slots (ms)=7872
                Total time spent by all map tasks (ms)=3764
                Total time spent by all reduce tasks (ms)=3936
                Total vcore-milliseconds taken by all map tasks=3764
                Total vcore-milliseconds taken by all reduce tasks=3936
                Total megabyte-milliseconds taken by all map tasks=963584
                Total megabyte-milliseconds taken by all reduce tasks=1007616
        Map-Reduce Framework
                Map input records=4
                Map output records=4
                Map output bytes=101
                Map output materialized bytes=115
                Input split bytes=129
                Combine input records=0
                Combine output records=0
                Reduce input groups=2
                Reduce shuffle bytes=115
                Reduce input records=4
                Reduce output records=4
                Spilled Records=8
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=78
                CPU time spent (ms)=890
                Physical memory (bytes) snapshot=358117376
                Virtual memory (bytes) snapshot=4097929216
                Total committed heap usage (bytes)=187980896
                Peak Map Physical memory (bytes)=227360768
                Peak Map Virtual memory (bytes)=2043523072
                Peak Reduce Physical memory (bytes)=130756608
                Peak Reduce Virtual memory (bytes)=2054406144
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=219
        File Output Format Counters
                Bytes Written=77
[hadoop@master ~]$
```

Verify Experiment 2

```
[hadoop@master ~]$ rm -fr output/
[hadoop@master ~]$ hdfs dfs -get output output
2021-12-10 11:08:49,546 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
lasses where applicable
[hadoop@master ~]$ cat output/*
2       dfs.replication
2       dfs.namenode.name.dir
2       dfs.datanode.data.dir
1       dfsadmin
[hadoop@master ~]$
```

Experiment 3 (Word Count ps.txt, top.txt, vi.txt)

I modified mapper.py, adding the regular expression operations library to remove all numbers and special characters from the output of the word count application. Highlighted are the changes I made to mapper.py

```
[hadoop@master ~]$ cat mapper.py
#!/usr/bin/env python
import sys
import re

#--- get all lines from stdin ---
for line in sys.stdin:

    #--- remove leading and trailing whitespace---
    line = line.strip()

    #--- remove special characters---
    line_wo = re.sub(r"[^a-zA-Z]", " ",line)

    #--- split the line into words ---
    words = line_wo.split()

    #--- output tuples [word, 1] in tab-delimited format---
    for word in words:
        print '%s\t%s' % (word, "1")
[hadoop@master ~]$
```

After the script removes the leading and trailing whitespace, I read that line eliminating the numbers and special characters before splitting the line into separate words.

Running experiment 3

```
[hadoop@master ~]$ mapred streaming -input /user/hadoop/txtInput/ -output txtOutput -mapper mapper.py -reducer reducer.py -file ma
pper.py -file reducer.py
2021-12-10 11:52:35,054 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
2021-12-10 11:52:35,250 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
packageJobJar: [mapper.py, reducer.py] [/home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.0.jar] /tmp/streamjob57585
8838975079381.jar tmpDir=null
2021-12-10 11:52:36,400 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/10.0.0.120:8032
2021-12-10 11:52:36,598 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/10.0.0.120:8032
2021-12-10 11:52:37,038 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.st
aging/job_1639149300891_0006
2021-12-10 11:52:38,082 INFO mapred.FileInputFormat: Total input files to process : 3
2021-12-10 11:52:38,544 INFO mapreduce.JobSubmitter: number of splits:4
2021-12-10 11:52:40,497 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1639149300891_0006
2021-12-10 11:52:40,497 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-12-10 11:52:40,750 INFO conf.Configuration: resource-types.xml not found
2021-12-10 11:52:40,750 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-12-10 11:52:40,811 INFO impl.YarnClientImpl: Submitted application application_1639149300891_0006
2021-12-10 11:52:40,897 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1639149300891_0006/
2021-12-10 11:52:40,898 INFO mapreduce.Job: Running job: job_1639149300891_0006
2021-12-10 11:52:48,153 INFO mapreduce.Job: Job job_1639149300891_0006 running in uber mode : false
2021-12-10 11:52:48,154 INFO mapreduce.Job:  map 0% reduce 0%
```

```
                Input split bytes=394
                Combine input records=0
                Combine output records=0
                Reduce input groups=3091
                Reduce shuffle bytes=210211
                Reduce input records=21927
                Reduce output records=3091
                Spilled Records=43854
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=567
                CPU time spent (ms)=3850
                Physical memory (bytes) snapshot=1032086504
                Virtual memory (bytes) snapshot=10244235264
                Total committed heap usage (bytes)=629338112
                Peak Map Physical memory (bytes)=229384192
                Peak Map Virtual memory (bytes)=2048651264
                Peak Reduce Physical memory (bytes)=120049664
                Peak Reduce Virtual memory (bytes)=2051792896
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=164194
        File Output Format Counters
                Bytes Written=28741
2021-12-10 11:53:15,681 INFO streaming.StreamJob: Output directory: txtOutput
[hadoop@master ~]$
```

Displaying the output from experiment 3

```
[hadoop@master ~]$ hdfs dfs -cat txtOutput/part-00000
2021-12-10 11:55:01,865 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
limited 4
searchable      1
represent       1
dynamic 1
sco     1
#GID    3
ENVIRONMENT     1
winname 1
collating       1
sleep   4
Fractional      2
consists        4
fsgid   1
pids    4
saved   8
aliases 2
whose   12
violate 1
ttys    1
swap    4
Additionally    2
bugreport       1
aux     6
spec    2
updated 1
unscrolled      1
ration  1
Debugging       1
XPO     1
```

I downloaded a book in text format and ran the word count application against it.

https://ia600908.us.archive.org/6/items/alicesadventures19033gut/19033.txt

```
[hadoop@master ~]$ mapred streaming -input /user/hadoop/assignment4/ -output bookOutput -mapper mapper.py -reducer reducer.py -fil
e mapper.py -file reducer.py
2021-12-10 14:31:20,225 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
2021-12-10 14:31:20,415 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
packageJobJar: [mapper.py, reducer.py] [/home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.0.jar] /tmp/streamjob41350
6406715508626.jar tmpDir=null
2021-12-10 14:31:21,617 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/10.0.0.120:8032
2021-12-10 14:31:21,834 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/10.0.0.120:8032
2021-12-10 14:31:22,547 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.st
aging/job_1639149300891_0007
2021-12-10 14:31:23,549 INFO mapred.FileInputFormat: Total input files to process : 1
2021-12-10 14:31:24,012 INFO mapreduce.JobSubmitter: number of splits:2
2021-12-10 14:31:24,569 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1639149300891_0007
2021-12-10 14:31:24,569 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-12-10 14:31:24,743 INFO conf.Configuration: resource-types.xml not found
2021-12-10 14:31:24,770 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-12-10 14:31:24,828 INFO impl.YarnClientImpl: Submitted application application_1639149300891_0007
2021-12-10 14:31:24,865 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1639149300891_0007/
2021-12-10 14:31:24,866 INFO mapreduce.Job: Running job: job_1639149300891_0007
2021-12-10 14:31:32,020 INFO mapreduce.Job: Job job_1639149300891_0007 running in uber mode : false
2021-12-10 14:31:32,021 INFO mapreduce.Job:  map 0% reduce 0%
2021-12-10 14:31:42,149 INFO mapreduce.Job:  map 100% reduce 0%
2021-12-10 14:31:50,227 INFO mapreduce.Job:  map 100% reduce 100%
2021-12-10 14:31:51,276 INFO mapreduce.Job: Job job_1639149300891_0007 completed successfully
2021-12-10 14:31:51,385 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=121047
                FILE: Number of bytes written=1044182
```

```
                Combine output records=0
                Reduce input groups=2282
                Reduce shuffle bytes=121053
                Reduce input records=13160
                Reduce output records=2282
                Spilled Records=26320
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=158
                CPU time spent (ms)=1810
                Physical memory (bytes) snapshot=590204928
                Virtual memory (bytes) snapshot=6147575808
                Total committed heap usage (bytes)=335179776
                Peak Map Physical memory (bytes)=232013824
                Peak Map Virtual memory (bytes)=2048081920
                Peak Reduce Physical memory (bytes)=131846144
                Peak Reduce Virtual memory (bytes)=2053226496
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=78822
        File Output Format Counters
                Bytes Written=20889
2021-12-10 14:31:51,385 INFO streaming.StreamJob: Output directory: bookOutput
```

Displaying the output for the word count application for the book Alice in Wonderland.

```
[hadoop@master ~]$ hdfs dfs -cat bookOutput/part-00000
2021-12-10 14:53:51,836 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
limited 2
pardon  3
secondly        1
saves   1
knelt   1
four    1
sleep   4
hanging 1
hate    2
assembled       2
forget  1
Foundation      24
calculate       1
swam    3
whatsoever      2
under   15
inwards 1
worth   1
void    1
rise    1
every   3
govern  1
bringing        1
MAD     1
leaders 1
tired   4
feathers        1
```

References

[1] *Talend.com*, 2021. [Online]. Available: https://www.talend.com/resources/what-is-mapreduce/. [Accessed: 10- Dec- 2021].

[2] *Talend.com*, 2021. [Online]. Available: https://www.talend.com/resources/what-is-hadoop/. [Accessed: 10- Dec- 2021].

[3]"re — Regular expression operations — Python 3.10.1 documentation", *Docs.python.org*, 2021. [Online]. Available: https://docs.python.org/3/library/re.html. [Accessed: 10- Dec- 2021].