John-Michael Woodrow CP1370 Lab 1

Part 1 mkdir

```
jmwoody133@cluster-2da1-m:~$ hadoop fs -mkdir /samplev2
jmwoody133@cluster-2da1-m:~$ hadoop fs -ls /
Found 8 items
drwxr-xr-x - jmwoody133 hadoop
                                       0 2025-01-17 16:10 /Input
-rw-r--r-- 2 jmwoody133 hadoop
                                      0 2025-01-17 16:33 /practice.txt
-rw-r--r-- 2 jmwoody133 hadoop
                                       0 2025-01-17 16:53 /practice2.txt
drwxr-xr-x - jmwoody133 hadoop
                                      0 2025-01-17 16:27 /sample
drwxr-xr-x - jmwoody133 hadoop
                                      0 2025-01-17 16:57 /samplev2
drwxrwxrwt - hdfs
                                      0 2025-01-17 15:51 /tmp
                      hadoop
drwxrwxrwt - hdfs
                        hadoop
                                      0 2025-01-17 15:51 /user
drwxrwxrwt
            - hdfs
                                      0 2025-01-17 15:51 /var
                        hadoop
jmwoody133@cluster-2da1-m:~$
```

Touch

```
jmwoody133@cluster-2da1-m:~$ hadoop fs -touch /practice3.txt
jmwoody133@cluster-2da1-m:~$ hadoop fs -ls /
Found 9 items
drwxr-xr-x - jmwoody133 hadoop
                                        0 2025-01-17 16:10 /Input
-rw-r--r- 2 jmwoody133 hadoop
                                         0 2025-01-17 16:33 /practice.txt
-rw-r--r-- 2 jmwoody133 hadoop
-rw-r--r-- 2 jmwoody133 hadoop
                                        0 2025-01-17 16:53 /practice2.txt
                                        0 2025-01-17 16:59 /practice3.txt
drwxr-xr-x - jmwoody133 hadoop
                                        0 2025-01-17 16:27 /sample
drwxr-xr-x - jmwoody133 hadoop
                                        0 2025-01-17 16:57 /samplev2
drwxrwxrwt - hdfs
                        hadoop
                                        0 2025-01-17 15:51 /tmp
                                        0 2025-01-17 15:51 /user
drwxrwxrwt - hdfs
                         hadoop
drwxrwxrwt - hdfs
                                        0 2025-01-17 15:51 /var
                         hadoop
jmwoody133@cluster-2da1-m:~$
```

appendToFile

```
jmwoody133@cluster-2da1-m:~$ echo "Hello World, this is a test, I'm not having fun" > localfile2.txt
jmwoody133@cluster-2da1-m:~$ hadoop fs -appendToFile localfile2.txt /practice3.txt
jmwoody133@cluster-2da1-m:~$ hadoop fs -cat /practice2.txt
jmwoody133@cluster-2da1-m:~$ hadoop fs -cat /practice3.txt
Hello World, this is a test, I'm not having fun
jmwoody133@cluster-2da1-m:~$
```

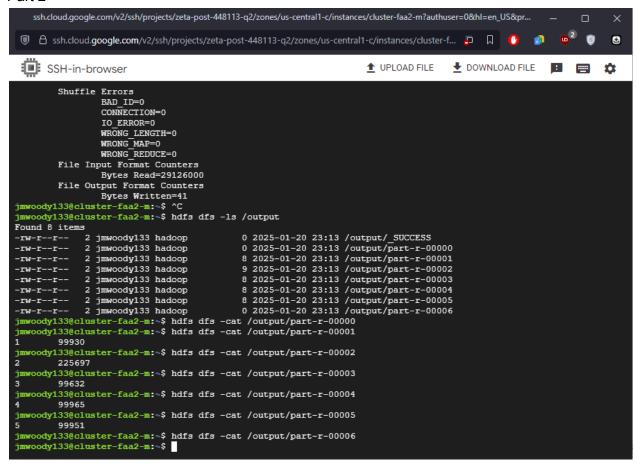
Count

```
jmwoody133@cluster-2da1-m:~$ hadoop fs -count /practice3.txt
0 1 48 /practice3.txt
jmwoody133@cluster-2da1-m:~$
```

Head

```
jmwoody133@cluster-2da1-m:~$ hadoop fs -cat /practice3.txt | head -n 1
Hello World, this is a test, I'm not having fun
jmwoody133@cluster-2da1-m:~$
```

Part 2



Part 3

- Big Data is a set of data that displays the characteristics of volume, velocity and variety to an extent that makes the data unsuitable for management by a relational database management system.
- 2. Volume Quantity of data to be stored.
 - Velocity Speed at which data is entering the system.
 - Variety Variations in the structure of the data to be stored.
- 3. Companies like Google and Amazon were among the first to address the Big Data Problem due to their need to process and analyze large amounts of data generated by their systems. These were some of the earliest companies to get massive amounts of traffic and in a way were forced to adapt to the high traffic.
- 4. Scaling up is keeping the same number of systems but migrating each to a larger system while scaling out is when the workload exceeds the capacity of a server, the workload is spread out across a number of servers, also referred to as clustering.

- 5. Stream processing is focusing on input processing and requires analysis of the data stream as it enters the system. It is necessary in some situations due to large columns of data entering the system at a fast pace that isn't feasible to store all the data. The data must be processed and filtered as it enters to determine what should be kept and what should be discarded.
- 6. Stream processing is different from feedback loop processing due to being thought of as focused on inputs while feedback loop processing can be thought of as focused on outputs.