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       hadoop          0 2025-01-17 15:51 /tmp
drwxrwxrwt   - hdfs       hadoop          0 2025-01-17 15:51 /user
drwxrwxrwt   - hdfs       hadoop          0 2025-01-17 15:51 /var
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          0 2025-01-17 16:53 /practice2.txt
-rw-r--r--   2 jmwoody133 hadoop          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
drwxrwxrwt   - hdfs       hadoop          0 2025-01-17 15:51 /user
drwxrwxrwt   - hdfs       hadoop          0 2025-01-17 15:51 /var
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

```
ssh.cloud.google.com/v2/ssh/projects/zeta-post-448113-q2/zones/us-central1-c/instances/cluster-faa2-m?authuser=0&hl=en_US&pr...
ssh.cloud.google.com/v2/ssh/projects/zeta-post-448113-q2/zones/us-central1-c/instances/cluster-f...
SSH-in-browser                                                    UPLOAD FILE    DOWNLOAD FILE

        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=29126000
        File Output Format Counters
                Bytes Written=41
jmwoody133@cluster-faa2-m:~$ ^C
jmwoody133@cluster-faa2-m:~$ hdfs dfs -ls /output
Found 8 items
-rw-r--r--   2 jmwoody133 hadoop          0 2025-01-20 23:13 /output/_SUCCESS
-rw-r--r--   2 jmwoody133 hadoop          0 2025-01-20 23:13 /output/part-r-00000
-rw-r--r--   2 jmwoody133 hadoop          8 2025-01-20 23:13 /output/part-r-00001
-rw-r--r--   2 jmwoody133 hadoop          9 2025-01-20 23:13 /output/part-r-00002
-rw-r--r--   2 jmwoody133 hadoop          8 2025-01-20 23:13 /output/part-r-00003
-rw-r--r--   2 jmwoody133 hadoop          8 2025-01-20 23:13 /output/part-r-00004
-rw-r--r--   2 jmwoody133 hadoop          8 2025-01-20 23:13 /output/part-r-00005
-rw-r--r--   2 jmwoody133 hadoop          0 2025-01-20 23:13 /output/part-r-00006
jmwoody133@cluster-faa2-m:~$ hdfs dfs -cat /output/part-r-00000
jmwoody133@cluster-faa2-m:~$ hdfs dfs -cat /output/part-r-00001
1       99930
jmwoody133@cluster-faa2-m:~$ hdfs dfs -cat /output/part-r-00002
2       225697
jmwoody133@cluster-faa2-m:~$ hdfs dfs -cat /output/part-r-00003
3       99632
jmwoody133@cluster-faa2-m:~$ hdfs dfs -cat /output/part-r-00004
4       99965
jmwoody133@cluster-faa2-m:~$ hdfs dfs -cat /output/part-r-00005
5       99951
jmwoody133@cluster-faa2-m:~$ hdfs dfs -cat /output/part-r-00006
jmwoody133@cluster-faa2-m:~$
```

Part 3

1. 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.