Hadoop

To use Hadoop, we **started the Hadoop services**:

```
$HADOOP_HOME/sbin/start-dfs.sh
$HADOOP_HOME/sbin/start-yarn.sh
```

We confirmed everything was running by checking:
```
jps
```

Expected running services:
```
NameNode
DataNode
ResourceManager
NodeManager
JobHistoryServer
```

Since Hadoop works on **HDFS**, we uploaded our input file:
hdfs dfs -put data.txt /

```
pyenv activate base
```

Ran the Hadoop streaming job:
```
hadoop jar /opt/homebrew/opt/hadoop/libexec/share/hadoop/tools/lib/hadoop-streaming-3.4.1.jar \
  -files count0.py \
  -input hdfs:///user/chkapsalis/data.txt \
  -output hdfs:///user/chkapsalis/output \
  -mapper "python3 count0.py --step-num=0 --mapper" \
  -reducer "python3 count0.py --step-num=0 --reducer"
```

Error Handling:
- **“File does not exist" Error**: Solved by ensuring count0.py was in the correct **local directory** (not HDFS).
- **“Output directory already exists" Error**: Solved by **deleting the existing output directory** before running a new job:
```
```

hdfs dfs -rm -r /user/chkapsalis/output. **# RUN THIS PRIOR TO RUNNING ANY NEW JOBS !!!**
```

### Once the job was successful, we checked the output:
```
hdfs dfs -ls /user/chkapsalis/output
hdfs dfs -cat /user/chkapsalis/output/part-00000
```

***

We encountered several issues when trying to restart Hadoop after a system reboot. Below is a list of all the problems you faced and the solutions I provided.
**## Problem 1: NameNode Not Starting**
**ERROR: Cannot set priority of namenode process**

## Cause:

- macOS **does not allow priority adjustments** for non-root processes.
- The **Hadoop NameNode storage directory was missing or deleted** due to a system reboot.

## Solution:

1. **Disable Priority Adjustment in Hadoop**
    - Edit Hadoop's environment file:
    - nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
    - Comment out “export HADOOP_NICENESS=0”
2. Recreate Missing NameNode Storage Directory
    1. run
        1. mkdir -p /private/tmp/hadoop-chkapsalis/dfs/name
        2. chmod 700 /private/tmp/hadoop-chkapsalis/dfs/name
    2. Restart hadoop
        1. $HADOOP_HOME/sbin/stop-dfs.sh
        2. $HADOOP_HOME/sbin/start-dfs.sh
    3. Reformat the namenode (if needed)
        1. hdfs namenode -format
    4. Change NameNode Storage to a Permanent Location
        1. nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
        2. ``` <property>  <name>dfs.namenode.name.dir</name> <value>/Users/chkapsalis/hadoop/dfs/name</value> </property> ```
    5. Create the new directory
        1. ``` mkdir -p /Users/chkapsalis/hadoop/dfs/name
        2. chmod 700 /Users/chkapsalis/hadoop/dfs/name ```
```

Problem 2: Datanodes not starting
java.io.IOException: All specified directories have failed to load.

Cause:

- The **DataNode storage directory was missing or inaccessible** after the system reboot.
- Hadoop **was still pointing to /private/tmp/**, which macOS deletes on restart.

Solution:

1. **Recreate the Missing DataNode Directory**
    1. Run
        1. ``` mkdir -p /private/tmp/hadoop-chkapsalis/dfs/data
        2. chmod 700 /private/tmp/hadoop-chkapsalis/dfs/data ```
    2. Restart Hadoop:
        1. ``` $HADOOP_HOME/sbin/stop-dfs.sh
        2. $HADOOP_HOME/sbin/start-dfs.sh ```
    3. Reformat the DataNode (if needed)
        1. hdfs datanode -format
    4. Move DataNode Storage to a Permanent Location
        1. Edit **hdfs-site.xml**
            1. nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
            2. ``` <property> <name>dfs.datanode.data.dir</name> <value>/Users/chkapsalis/hadoop/dfs/data</value> </property> ```
    5. Create the new directoy
        1. ``` mkdir -p /Users/chkapsalis/hadoop/dfs/data
        2. chmod 700 /Users/chkapsalis/hadoop/dfs/data ```

Problem 3: NameNode Couldn’t Connect to DataNode
java.io.IOException: Initialization failed for Block pool <registering>

Cause:

- The **DataNode was not properly formatted** or **not able to communicate** with the NameNode.

Solution:

1. **Ensure the DataNode directory is accessible**
    - Check if it exists:
        - ``` ls -ld /Users/chkapsalis/hadoop/dfs/data ```
    - If not, recreate it:
        - ``` mkdir -p /Users/chkapsalis/hadoop/dfs/data
        - chmod 700 /Users/chkapsalis/hadoop/dfs/data ```
- Reformat and Restart Hadoop Services
        - ``` hdfs datanode -format
        - $HADOOP_HOME/sbin/stop-all.sh

- hdfs namenode -format # Only if necessary!
- hdfs datanode -format # Only if necessary!
- $HADOOP_HOME/sbin/start-dfs.sh
- $HADOOP_HOME/sbin/start-yarn.sh ```