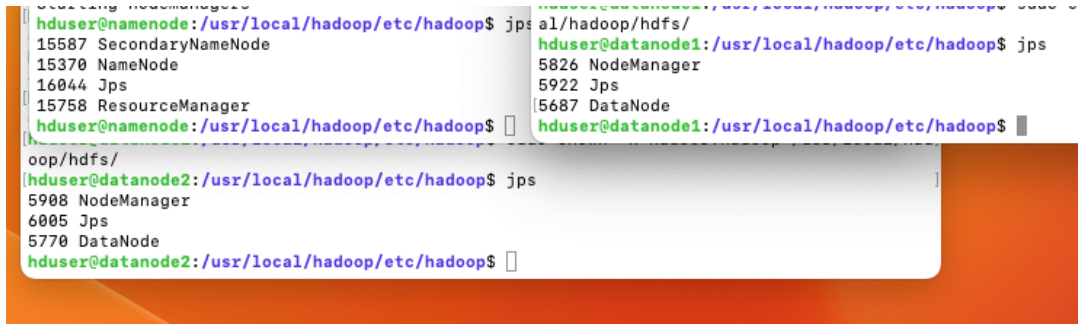


## Lab2

Steps:

1. Set up a 3 node cluster with Hadoop



```

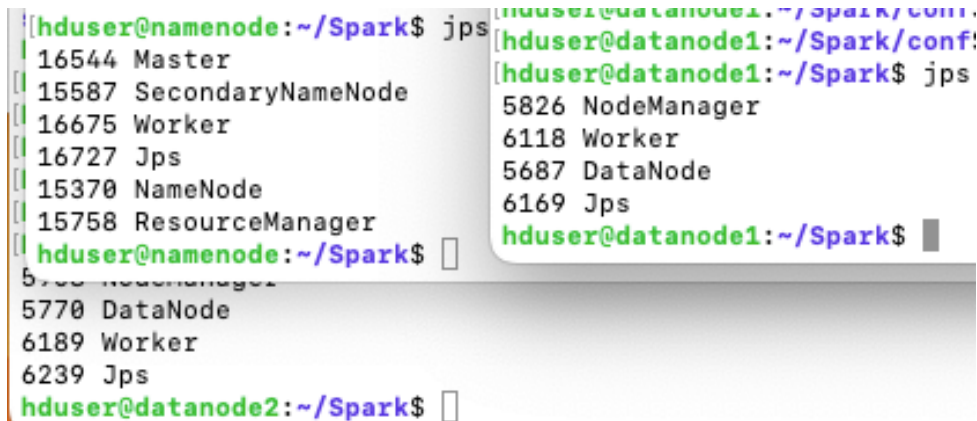
hduser@namenode: /usr/local/hadoop/etc/hadoop$ jps
15587 SecondaryNameNode
15370 NameNode
16044 Jps
15758 ResourceManager
hduser@namenode: /usr/local/hadoop/etc/hadoop$

hduser@datanode1: /usr/local/hadoop/etc/hadoop$ jps
5826 NodeManager
5922 Jps
5687 DataNode
hduser@datanode1: /usr/local/hadoop/etc/hadoop$

hduser@datanode2: /usr/local/hadoop/etc/hadoop$ jps
5908 NodeManager
6005 Jps
5770 DataNode
hduser@datanode2: /usr/local/hadoop/etc/hadoop$

```

2. Install Spark and configure- run Spark



```

hduser@namenode: ~/Spark$ jps
16544 Master
15587 SecondaryNameNode
16675 Worker
16727 Jps
15370 NameNode
15758 ResourceManager
hduser@namenode: ~/Spark$

hduser@datanode1: ~/Spark/conf$ jps
5826 NodeManager
6118 Worker
5687 DataNode
6169 Jps
hduser@datanode1: ~/Spark$

hduser@datanode2: ~/Spark$ jps
5770 DataNode
6189 Worker
6239 Jps
hduser@datanode2: ~/Spark$

```

3. Upload .csv files and store in HDFS  
`hduser@namenode: ~$ hdfs dfs -put /home/hduser/shot_logs.csv /input`  
`hdfs dfs -put /home/hduser/Parking_Violations_Issued_-_Fiscal_Year_2023.csv /input`
4. Create new conda environment:  
`conda create -n pyspark python=3.8`  
`conda activate pyspark`  
`conda install pyspark`
5. Create Jupyter Notebook  
`conda install jupyter`  
`jupyter notebook`
6. At this point I started running into errors and I had already spent ~20 hours trying to get the cluster to work (ran into issues cloning the repository as well so went with Pyspark), so I decided to write the code with RDD transformations without being able to check. Please see accompanying ipynb files.
7. Future steps: I believe some of the errors were related to Java gateway process being exited, and the need for configuration of HADOOP\_CONF\_DIR or YARN\_CONF\_DIR.