**\*\*\*SCREENSHOTS OF OUTPUT FOR EACH PROGRAMMING PROBLEM ARE AT THE END OF THIS DOCUMENT\*\*\***

**CPSC 4770: Assignment 4**

**Reagan Leonard & Ethan McGaha**

**Environment Setup commands**

cd myhadoop

./init\_hadoop.sh

export HADOOP\_CONF\_DIR="/home/$USER/hadoop\_palmetto/config/"

hdfs dfs -mkdir /user/

hdfs dfs -mkdir /user/$USER

hdfs dfs -mkdir /user/$USER/intro-to-hadoop

hdfs dfs -put /zfs/citi/movielens intro-to-hadoop/

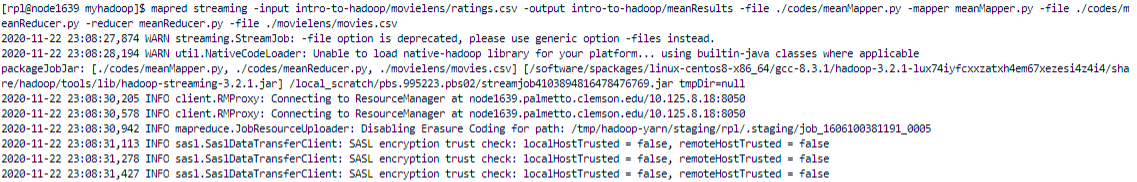
**Mean commands**

**Non-Streaming**

hdfs dfs -cat intro-to-hadoop/movielens/ratings.csv 2>/dev/null | head -n 10 | python ./codes/meanMapper.py | sort | python ./codes/meanReducer.py

**Streaming**

mapred streaming -input intro-to-hadoop/movielens/ratings.csv -output intro-to-hadoop/meanResults -file ./codes/meanMapper.py -mapper meanMapper.py -file ./codes/meanReducer.py -reducer meanReducer.py -file ./movielens/movies.csv





hdfs dfs -ls intro-to-hadoop/meanResults

hdfs dfs -cat intro-to-hadoop/meanResults/part-00000 2>/dev/null | head -n 20

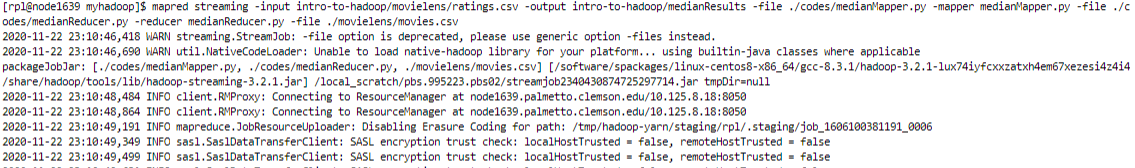
**Median commands**

**Non-Streaming**

hdfs dfs -cat intro-to-hadoop/movielens/ratings.csv 2>/dev/null | head -n 10 | python ./codes/medianMapper.py | sort | python ./codes/medianReducer.py

**Streaming**

mapred streaming -input intro-to-hadoop/movielens/ratings.csv -output intro-to-hadoop/medianResults -file ./codes/medianMapper.py -mapper medianMapper.py -file ./codes/medianReducer.py -reducer medianReducer.py -file ./movielens/movies.csv





hdfs dfs -ls intro-to-hadoop/medianResults

hdfs dfs -cat intro-to-hadoop/medianResults/part-00000 2>/dev/null | head -n 20

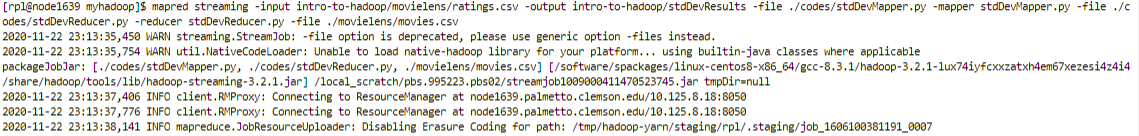
**Standard Deviation commands**

**Non-Streaming**

hdfs dfs -cat intro-to-hadoop/movielens/ratings.csv 2>/dev/null | head -n 10 | python ./codes/stdDevMapper.py | sort | python ./codes/stdDevReducer.py

**Streaming**

mapred streaming -input intro-to-hadoop/movielens/ratings.csv -output intro-to-hadoop/stdDevResults -file ./codes/stdDevMapper.py -mapper stdDevMapper.py -file ./codes/stdDevReducer.py -reducer stdDevReducer.py -file ./movielens/movies.csv





hdfs dfs -ls intro-to-hadoop/stdDevResults

hdfs dfs -cat intro-to-hadoop/stdDevResults/part-00000 2>/dev/null | head -n 20

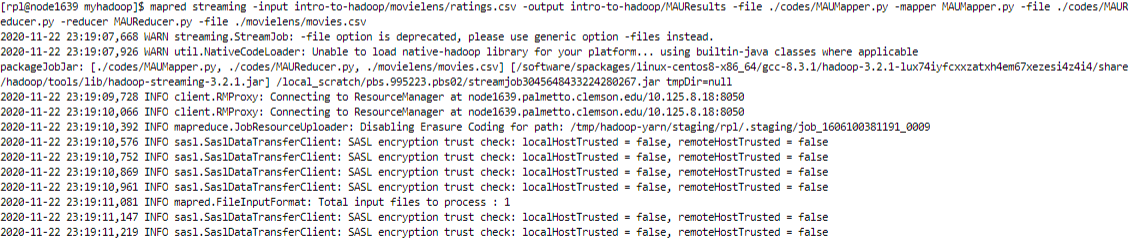
**Most Active User commands**

**Non-Streaming**

hdfs dfs -cat intro-to-hadoop/movielens/ratings.csv 2>/dev/null | head -n 10 | python ./codes/MAUMapper.py | sort | python ./codes/MAUReducer.py

**Streaming**

mapred streaming -input intro-to-hadoop/movielens/ratings.csv -output intro-to-hadoop/MAUResults -file ./codes/MAUMapper.py -mapper MAUMapper.py -file ./codes/MAUReducer.py -reducer MAUReducer.py -file ./movielens/movies.csv





hdfs dfs -ls intro-to-hadoop/MAUResults

hdfs dfs -cat intro-to-hadoop/MAUResults/part-00000 2>/dev/null | head -n 20

**Test Streaming commands (slide 13 on Slides\_14)**

mapred streaming -input intro-to-hadoop/movielens/ratings.csv -output intro-to-hadoop/output-movielens-02 -file ./codes/avgRatingMapper04.py -mapper avgRatingMapper04.py -file ./codes/avgRatingReducer01.py -reducer avgRatingReducer01.py -file ./movielens/movies.csv

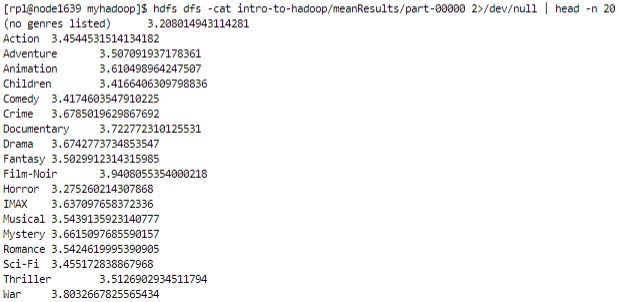
hdfs dfs -ls intro-to-hadoop/output-movielens-02

hdfs dfs -cat intro-to-hadoop/output-movielens-02/part-00000 2>/dev/null | head -n 20

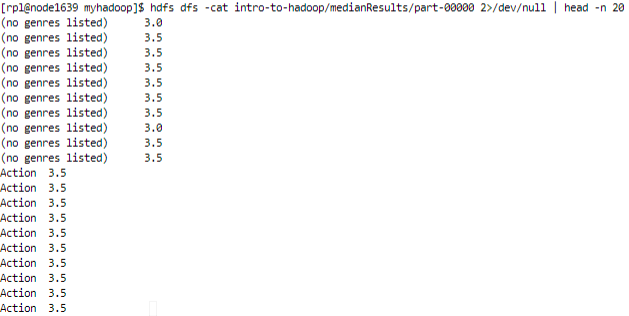
**Integrating into Palmetto Workflow**

* Open a terminal, ssh to login001 (DUO required), and submit this script
  + ssh login001
  + cd
  + cat –n ~/myhadoop/codes/movieAnalyzer.pbs
  + qsub ~/myhadooop/codes/movieAnalyzer.pbs
* View the final output when the job is finished
  + qstat –anu $USER
  + cat part-00000 2>/dev/null | head -n 20

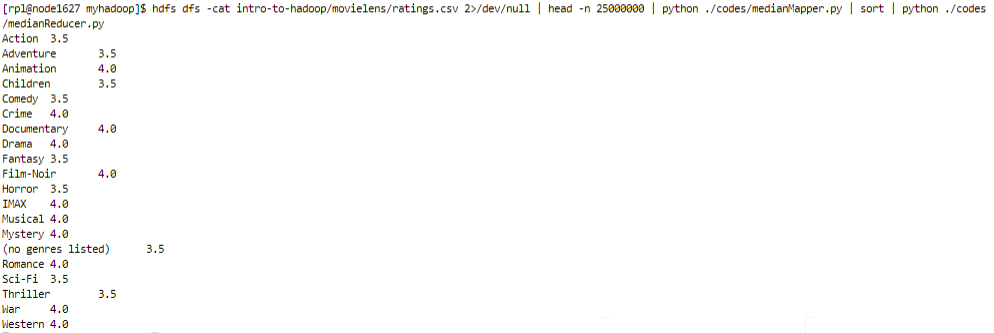
**Mean Results**



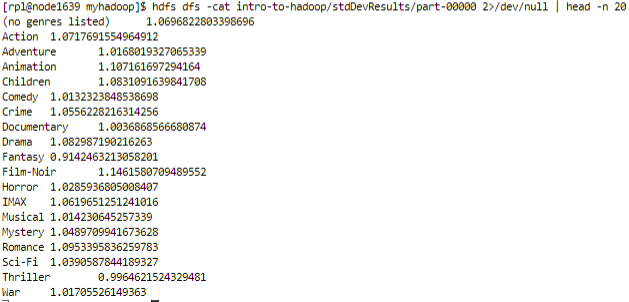
**Median Results**



The results above are formatted incorrectly. When we run the mapred streaming command, each genre’s median gets printed to the output file 10 times for some reason. The medians are still correct though. This formatting error does not happen, however, when we run the simpler, non-streaming hdfs command. So for convenience sake, we have included a screenshot of those results below as well. (This non-streaming command was run with all the ratings.)



**Standard Deviation Results**



**Most Active User Results**

