Simon Doku & Joshua Martinez

Dr. Hala ElAarag

CSCI301

24 April 2024

Project 3 Write Up

Dataset Description

For this project the “names corpus” dataset was used from Kaggle. This dataset contains two files, one with 5001 female names female names in alphabetical order nd the other with male names. For simplicity we decided to use just one of them- the female names dataset.

The program was written in python and was made to perform the following operations using Spark:

1. Word count: gives the total count of names in the file
2. Longest word: displays the longest word in the dataset
3. Occurrences of entire word: searches for any occurrence of the entire word in the dataset
4. Search for any occurrence of word: finds any occurrence of the word in the dataset
5. Time taken: Calculates time taken to complete the entire operation in seconds

Results

Results remain undefined. Despite having the program run, it would keep running. The longest run before manual release took 30 minutes with no output.

Analysis

The failure to complete the job is clearer with the challenges faced, but the idea is that one Raspberry Pi was not enough to parse data the size we requested. The only way to benefit from utilizing Spark with Raspberry Pis is to employ a supercluster with multiple units.

Challenges

The biggest challenges came from setting up Apache Spark. Some errors came up on every turn. Some relate the version of Java to the Spark being updated moments after setting up the master node. Additionally, the deprecation of the start-slaves script hadn't come up in the initial setup, so pivoting from the instructions made things difficult.

The biggest and most long-standing error came from launching our workers. Utilizing the recommended start-workers script didn't produce any results, so we needed to ssh into the individual nodes and run the start-worker script. This yielded an exception for creating a too large frame, which prompted a second and third reinstallation of Apache Spark to try different versions.

Online resources seemed to indicate a few possible avenues to explore. One of these was to get a larger capacity boot SD card. This yielded no results. Another potential avenue was configuring the firewall settings and ensuring the required ports were open. This did nothing to fix the error either. Another option was to double and triple-check the installation, which, once again, yielded no results.

The solution was tried before on a previous version of Spark, but that time it yielded no results. The error was in the command to start the worker. We used port 8080 instead of 7077. While this didn't work on version 3.4.2, it worked on 3.5.1. This only proved to be a temporary success. No additional workers would be allowed to join the master.

Conclusion

To spend nearly 20 hours working on a project like this to receive no results is quite upsetting. However, the proficiencies gained in navigating Linux, learning networking, and working with complex applications are valuable, not to mention the problem-solving skills gained by working incredibly hard to meet a deadline.

Under ideal conditions, if there was more time and capability to dedicate more time, a potential reinstall might be able to fix this. Eventually, our units were all clones of one worker, while the master was unique. This undoubtedly caused issues. A fresh OS install would be helpful here.