**Working with Big Data**

Chioma Chance

Colorado State University Global

CSC507: Ethical Leadership in Software Development

Dr. L

02/09/2025

**Efficient Processing of Large Files Using Parallel Processing Techniques**

**Introduction**  
The goal of this project was to apply the concepts learned in this course to efficiently process big data files by leveraging parallel processing techniques. This assignment required creating two programs to handle files containing large datasets, totaling one billion rows each, and output their sums into a new file, all while optimizing execution time. Although the original goal involved running multiple parallel processes, limitations in available storage and resources caused the assignment to stop after the second method, which processes two parts. Despite these challenges, this exercise provided valuable insights into handling large files in resource-constrained environments.

**Methodology**  
The assignment began by generating two large files, *hugefile1.txt* and *hugefile2.txt*, each containing 10 million lines of random numbers. Using a shell script, the task was to read and process these files concurrently, outputting their summed values into a final file, *totalfile.txt*. The process was divided into three parts:

1. **Full File Processing**: A single shell script (*process\_full.sh*) processed the entire files sequentially. The script used a combination of the paste command to merge lines from both input files and awk to calculate their sum. The elapsed time for this operation was approximately **354 seconds**, as captured in the screenshot provided.
2. **Processing Two Parts**: The next method divided both files into two parts using the split command. Each part was processed independently using *process\_2\_parts.sh*, a script designed to handle half of the files at a time. The results from the two parts were then concatenated into a single output file. However, this step was interrupted due to disk space limitations, so the runtime could not be recorded.

**Challenges Encountered**  
Throughout the project, several challenges impacted the results. The most significant issue was the lack of sufficient disk space within the virtual machine environment, which led to repeated interruptions. This limitation prevented completion of the third method, which involved dividing the files into ten parts for parallel processing. Additionally, generating large files caused significant strain on the system, with the terminal occasionally freezing or crashing.

**Results and Observations**  
The results obtained revealed key performance differences between methods:

* The full processing method demonstrated the simplicity of sequential execution, but its runtime highlighted the inefficiency of processing extremely large files without parallelism.
* Splitting the files into two parts was expected to halve the processing time, but this could not be verified due to resource constraints.

The screenshots provided evidence of the elapsed time for the full-file method and the output file contents. These constraints reinforced the importance of resource planning when dealing with big data tasks.

**Conclusion**  
This project underscored the value of parallel processing in handling large datasets. While it was not possible to complete all intended methods, the results demonstrated that even splitting files into two parts can improve processing efficiency if adequate resources are available. Moving forward, system configurations, including storage and memory, must be carefully considered to optimize performance for big data projects. The lessons learned from this experience will inform future approaches to working with resource-intensive tasks.

**Screenshots**A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

**References**

* Silberschatz, A., Galvin, P. B., & Gagne, G. (2018). *Operating system concepts* (10th ed.). Wiley.
* Bryant, R. E., & O'Hallaron, D. R. (2015). *Computer systems: A programmer's perspective* (3rd ed.). Pearson.
* INRIA. (2019). Proceedings of the 10th real-time systems open problems seminar. Retrieved from <https://project.inria.fr/rtsops2019/files/2019/07/Proceedings-RTSOPS2019.pdf>