Proyecto 01

Introducción: La implementación de procesamiento distribuido cobra auge en tareas que se centran en la utilización de grandes cantidades de recursos. Ciertas ocasiones no se tienen tareas altamente demandantes en cuanto a procesamiento (tareas complejas) sino son tareas repetitivas que, son pesadas debido a la cantidad de datos que deben de tomarse en cuenta. Archivos pesados para procesamiento de modelos matemáticos, análisis de datos basados en fuentes de datos no relacionales o de archivos de texto son algunos ejemplos de lo que se puede llegar a simplificar con el correcto uso de hilos o procesos.

Objetivos:

- Que el estudiante aplique los conceptos de creación de procesos a bajo nivel.
- Que el estudiante cree aplicaciones que tengan interfaz en terminal, así como también procesos e hilos en background.
- Que el usuario se familiarice con C y su uso a bajo nivel.
- Modificar estados de hilos (Crear, pausar, eliminar).
- Sincronizar threads.

Descripción: Se solicita una aplicación la cual sea capaz de contabilizar las diferentes letras que se encuentren dentro de un archivo separado por comas (csv). El objetivo es que a medida que va contabilizando, se pueda ir mostrando la cantidad de letras que va encontrando.

Por ejemplo, si el archivo tuviese 3 columnas, se viera de la siguiente manera:

Archivo

c1	c2	c3
A	D	G

Resultado

A	1
D	1
G	1

Imagen 01: Ejemplo archivo de entrada de 3 columnas y resultado esperado

Archivo

c1	c2	c3	c4	c5
A	D	A	B	D

Resultado

A	2
B	1
D	2

Imagen 02: Ejemplo de archivo de entrada de 5 columnas y resultado esperado

Aunque el archivo debe de ir mostrando el resultado de lo contabilizado en tiempo real, al terminar la cuenta, deberá de almacenar este resultado en un archivo csv de 2 columnas (una para la letra y otra para la cantidad registrada).

Requisitos técnicos:

- La aplicación deberá de ser creada en C.
- La máquina virtual en donde funcionará la aplicación deberá de tener como máximo 1GB de RAM y 2 núcleos de procesamiento sobre un sistema operativo Linux.
- La aplicación deberá de soportar archivos csv de cualquier tamaño. El requisito básico es que el archivo de entrada tenga columnas nombradas c1, c2, c3, cN y que los datos contenidos sean 1 solo carácter. Este carácter serán las letras de la A a la Z (no importa que sean mayúsculas o minúsculas, ya que serán únicamente para contar los caracteres). Nota: el nombre de las columnas puede cambiar por lo que la programación de la aplicación no debe de basarse en nombre de las columnas sino por posición.
- Debe de mostrar el tiempo que lleva en ejecución la tarea, recursos utilizados tanto en RAM como en procesador. La forma de mostrar estos valores queda a discreción del equipo de trabajo, pero es requisito para la calificación.

Forma de calificación/Presentación

Para validar el funcionamiento de la aplicación, se pedirá al grupo que carguen archivos de diferentes tamaños.

- La primera prueba será un archivo con 1 sola fila en donde se sepa a simple vista la cantidad de letras que contiene. Se cargará el archivo a la aplicación y se esperará a ver el resultado.
- La segunda prueba será nuevamente un archivo pequeño, esta vez de 200 columnas y 1 fila. Se validará el resultado que sea correcto.
- Posterior a estas 2 pruebas básicas, se solicitará la carga de un archivo el cual tenga 200 columnas y que tenga un peso de 2.5GB. La cantidad de filas dependerá de este peso, pero se estima que serán alrededor de 3.2 millones de filas en el documento.
- Esta prueba con el archivo de al menos 2.5GB debe de tardar no más de 20 minutos en ejecución.

MVP: Crear una aplicación que permita la contabilización de letras con un archivo de 2.5GB como entrada y devuelva un archivo csv con 2 columnas con el resultado. Dicha aplicación deberá de cumplir los requisitos técnicos anteriormente descritos.

Entregable: a diferencia de otras prácticas y laboratorios. Este proyecto tiene como objetivo incentivar la investigación por parte del equipo de trabajo. Se espera un documento formal de investigación de no más de 3 páginas las cuales describan el proceso por el cual su código funciona, es eficiente, divide de búsqueda en los recursos disponibles. El objetivo es la investigación del funcionamiento a detalle y a bajo nivel a manera que, si este documento se publicara, pudiera ser tomado como base para una nueva librería en C para búsqueda y contabilización de datos dentro de un archivo en un entorno de hardware limitado. La elaboración de este documento tiene como objetivo que sea de ayuda para entender maneras de optimizar su código previo a hacerlo y no hacer código sin realizar un análisis y diseño previos.

A fin de incentivar la búsqueda, investigación y exhortar un mejor desarrollo de proyecto la calificación del proyecto será dada por:

- 10 puntos – Documentación (anteriormente descrita)
- 70 puntos – Funcionamiento (MVP como mínimo para calificación)
- 20 puntos – Velocidad de algoritmo. Tomando en cuenta que todos deberán de tener un hardware idéntico (anteriormente descrito). Todos los proyectos deberán de estar listos a las 10 am del día 14 de marzo en el salón de clases. A esa hora, los proyectos se pondrán a ejecutar con el mismo archivo. Al equipo que termine de primero con la contabilización correcta, se le otorgarán sus 20 puntos. Quien termine de segundo, 18 puntos y así sucesivamente. Al ser 10 equipos, los últimos 2 equipos podrán optar únicamente a 2 de estos 20 puntos de “competencia”. **Nota**: los grupos que no estén listos a las 10 am para iniciar la competencia, no podrán optar a estos puntos.

Recomendaciones:

- Optimice el uso de threads en su aplicación.
- Tome en cuenta que múltiples threads no se vayan a intersectar el trabajo y así devuelvan contabilidades erróneas.
- Monitoree el uso de recursos de su máquina virtual, use de la mejor manera posible todo el entorno tanto de SO como de hardware que posea