UNIVERSIDAD NACIONAL EXPERIMENTAL DE GUAYANA

VICERRECTORADO ACADEMICO

COORDINACIÓN DE INGENIERÍA EN INFORMÁTICA

COORDINACIÓN GENERAL DE PREGRADO

MATERIA: SISTEMAS DISTRIBUIDOS

SEMESTRE VII

SECCION I

TAREA I

ALGORITMO MAPREDUCE

PROFESORA:

ING. VIRGINIA PADILLA

BACHILLERES:

ERIKA BRITO

GEORGE GUEVARA

CIUDAD GUAYANA, 20 DE MAYO DE 2024

INDICE

INTRODUCCION

Uno de los términos, comúnmente utilizado en la carrera de ingeniería en informática es el de sistemas y se puede definir como:

“Sistema es una entidad formada por un conjunto de elementos o componentes básicos del sistema, y por las relaciones existentes entre ellos, así como con el entorno. Estas relaciones se expresan formalmente empleando lenguaje matemático”. (Platero,2012)

Es decir, es un conjunto de partes compuestos por hardware y software unidos entre sí para cumplir como objetivo una tarea en específico. Los sistemas pueden clasificarse en centralizados y distribuidos, en el caso de la investigación realizada se fundamenta en los conceptos, principios, características y arquitecturas implementada en los sistemas distribuidos.

¿Qué es un sistema Distribuidos?

“Un sistema distribuido es una colección de computadoras independientes que dan al usuario la impresión de constituir un único sistema” (Tanenbaum, Van Steen, 2008).

¿Qué es un sistema centralizado?

Se puede decir que un sistemas centralizado es, “Donde una sola computadora ejecuta el software que proporciona un servicio, o la computación cliente-servidor, donde varias máquinas accesan de manera remota un servicio centralizado” (Padilla, 2024)

Partiendo de este concepto, se estudia el algoritmo MapReduce en un sistema centralizado y un sistema distribuido con la implementación de hilos (threads) para comparar el comportamiento de ambos sistemas y las ventajas que ofrece cada uno. Con el desarrollo del siguiente documento se conocerá la definición del algoritmo MapReduce, ventajas, uso, hilos como implementación de subprocesos dentro del software desarrollado y los resultados obtenidos en tiempo al correr ambos casos.

¿Qué es MapReduce?

Según Rivadeneira, 2023 MapReduce es “un modelo de programación en los procesos Hadoop, diseñado para la eficiencia y escalabilidad en entornos distribuidos. Reduce, implica la acumulación y combinación de datos para producir un resultado único, mientras que Map, se centra en aplicar una función a cada elemento de un conjunto de datos”.

¿Para qué sirve el algoritmo MapReduce?

Este algoritmo permite resolver tareas muy complejas, que generan alto costo en la ejecución de los procesos. Según (Casero, 2024) una de las aplicaciones que ha tenido el algoritmo es en BigData y ciencia de datos porque, “Te permite distribuir la carga de trabajo en múltiples nodos de un clúster de servidores, lo que acelera significativamente el procesamiento de grandes volúmenes de datos.”

¿En qué consiste el algoritmo MapReduce?

En base al enunciado de la tarea I de la asignatura sistemas distribuidos el algoritmo, permite la entrada de archivos agrupados e implementa dos funciones básicas como son:

- Se define una función map que transforma esos pares en otros diferentes donde la clave es una palabra concreta y el valor siempre es 1, es decir, genera un elemento “(palabra, 1)” por cada palabra que ve en los documentos que procesa.
- Se define una función reduce que se encarga de procesar todos los elementos “(palabra, 1)” de la misma palabra y calcular la suma final para cada palabra.

Herramientas, resolución de la tarea I Sistemas Distribuidos.

1. **Lenguaje de programación:** Python.
2. **Tipo de programación:** Programación estructurada.
3. **Implementación de librerías:**
    - Drive: Permite la conexión a google drive, donde se encuntra el archivo pagina.txt y se almacenaran los archivos Output.csv y conteo.csv.
    - Requests: Permite las solicitudes HTTP.
    - BeautifulSoup: Permite leer el código HTML de cada pagina web.
    - Re: Permite el manejo de caracteres.
    - Threading: Permite crear lo hilos como subprocesos de cada algoritmo
    - Queue: Permite el manejo de colas.

4. **Entorno de desarrollo:** Google Colab, se desarrolló en este entorno por las bondades que ofrece como acceso gratuito a GPUs, integración a GoogleDrive, integración a GitHub, escritura y ejecución de código en Python asi como la fácil importación de librerías Python.
5. **Código base y modelado de software:** Para la simulación del algoritmo en sistemas distribuidos, haciendo uso de hilo se utilizó como base el código de (Vipanchi Reddy Katthula, 2020).
6. **Documento de entrada:** Paginas.txt
7. **Documentos de Salida:** En el caso de un hilo nuevoconteo.csv y en el caso de multihilo Output.csv
8. **Documento de código para MapReduce en un hilo:** mapreduceone.ipynb
9. **Documento de código para MapReduce multihilo:** Actividad1 distribuido.ipynb

CONCLUSIONES

Concluida la investigación de tipo documental y el desarrollo de ambos algoritmo se puede decir, que el algoritmo MapReduce implementado en un hilo tardo más tiempo, que cuando se implementa en multihilos, tomando en cuenta que la cantidad de páginas web a revisar son 4 url, como muestra la siguiente tabla:

Sistema	Lenguaje	N° de hilo	Archivo Entrada	N° de paginas	Tiempo de Ejecución
Centralizado	Python	1	Paginas.txt	4	4s
Distribuido	Python	2	Paginas.txt	4	3s

Tabla 1. Propia Fuente

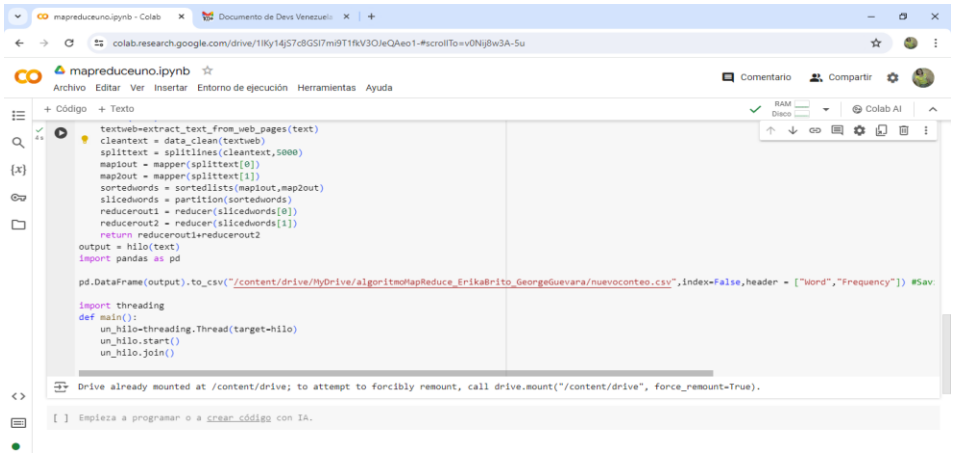

Imagen 1. Propia Fuente

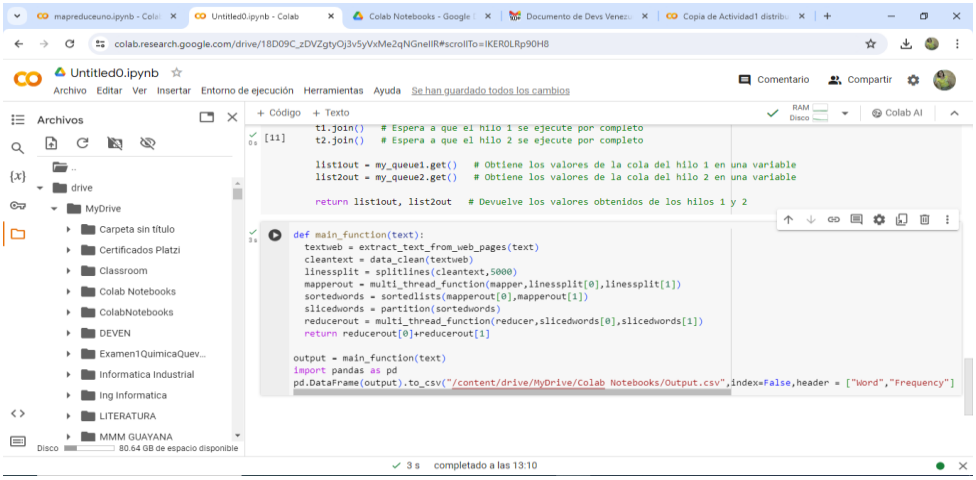

Imagen 2. Propia Fuente

REFERENCIAS

- Casero A, 2024. Que es MapReduce en Python. (Documento en línea) https://keepcoding.io/blog/que-es-mapreduce-en-python/
- Platero, C. Apuntes de Regulación Automática, 2012
- Padilla V, 2023. Introducción a sistemas distribuidos. Primera edición,2023.
- Rivadeneira L, 2023. MapReduce: Optimización en programación paralela y distribuida. (Documento en línea) https://medium.com/@liziel.rivadeneira.dso/mapreduce-optimizaci%C3%B3n-en-programaci%C3%B3n-paralela-y-distribuida-f9d33c036b8b
- Tanenbaum A, Van Steen M, 2008. Sistemas Distribuidos. Principios y Paradigmas. Segunda Edición. Editorial Prentice Hall, México, 2008.
- Vipanchi Reddy Katthula, 2020. Map Reduce 101-Python implementation (MultiThreadindg) (Documento en línea) https://vipanchikatthula.github.io/post/mapper-reducer-implementation/