OBSERVACIONES LABORATORIO 6

María Alejandra Moreno Bustillo, 202021603

Juliana Delgadillo Cheyne, 202020986

1. **Teniendo en cuenta cada uno de los requerimientos ¿Cuántos índices implementaría en el Reto? y ¿Por qué?**
   R// En total tendríamos 5 índices para el reto. Estos índices serían los siguientes: 1) Año de nacimiento del artista, 2) Año de adquisición de la obra, 3) Artistas con otro map asociado que clasifique sus obras por técnica, 4) Nacionalidad de los artistas de las obras y 5) Departamento al que pertenece la obra. Implementaríamos estos 5 índices puesto que facilitarían mucho más la búsqueda de datos para cumplir con los requerimientos de una manera mucho más eficiente que con el TAD lista en cuanto a tiempos de ejecución. Con la información ya clasificada según lo que piden los requerimientos, entregar las respuestas será mucho más rápido.

2. **Según los índices propuestos ¿en qué caso usaría Linear Probing o Separate Chaining en estos índices? y ¿Por qué?**

   R// Para los índices decidimos utilizar en todos ellos Linear Probing porque creemos que esta estructura nos permitirá lograr mejores tiempos en la ejecución de nuestras consultas. Sabemos que este necesitará de mayor espacio de almacenamiento en nuestro computador, sin embargo, dadas las características de la máquina que estamos utilizando, consideramos que será mejor emplear mayor memoria a cambio de menores tiempos de respuesta en los requerimientos. Por otro lado, a diferencia de Separate Chaining, esta estructura no encadena datos dentro de una misma posición, lo cual nos permitirá encontrar la información de una forma más sencilla y rápida. Sin embargo, esto hará que la carga de datos sea más demorada que en el caso de separate chaining, pero es un inconveniente mejor, pues al final de todo se busca que los tiempos de respuesta de los requerimientos sean lo menor posible, no el tiempo de la carga de datos.

3. **Dado el número de elementos de los archivos MoMA, ¿Cuál sería el factor de carga para estos índices según su mecanismo de colisión?**
   R// Para el archivo Artist - Large, el total de artistas encontrados fue de 15223 artistas, para el factor de carga esperado sería n/m donde n es la cantidad total de artistas sobre la cantidad total de espacios en la tabla de hash, en este caso se esperaría que ese m fuera el doble de n y se toma el número primo mayor más cercano, esto nos daría el siguiente valor para m: 30449. Esto nos daría un factor de carga de 0.499 que aproximado daría el factor de carga de 0.5. Asimismo, para el archivo Artwork- Large, el total de las obras encontradas es de 138150 obras, por esta razón el factor de carga será de n/m siendo n la cantidad total de obras y m el total de posiciones en la tabla hash, es decir 138150/276319 = 0.499 que se aproxima a 0.5. Estos resultados serían en el caso de que se haga uso de linear probbing, con separate chaining los factores de carga suelen ser mayores y, por ende, la m suele ser menor, pero todo depende del factor de carga que se escoja.

4. **¿Qué diferencias en el tiempo de ejecución notan al ejecutar la cargar los datos al cambiar la configuración de Linear Probing a Separate Chaining?**

tiempo_chaining_4 = 81687.5

tiempo_probing_0.5 = 88734.375

Luego de ejecutar la carga de datos con las diferentes configuraciones, se encuentra que al usar Separate Chaining (factor de carga de 4.00) la ejecución toma un tiempo de 81687.5ms, mientras que con Lineal Probing (factor de carga de 0.5) se demora 88734.375ms. Como es evidente, Probing aumenta el tiempo 7046.875ms en comparación con la primera configuración, por lo cual la opción más eficiente en cuanto a tiempos de carga será utilizar Separete Chaining con factor de carga de 4.00, para realizar la carga de datos de nuestro programa.

Se realizan pruebas de tiempo para Separate Chaining y Lineal Probing con distintos factores de carga como se muestra a continuación.

tiempo_chaining_2 = 86546.875

tiempo_chainig_8 = 81156.25

tiempo_probing_0.2 = 84968.75

tiempo_probing_0.8 = 93812.5

5. **¿Qué configuración de ADT Map escogería para el índice de técnicas o medios?, especifique el mecanismo de colisión, el factor de carga y el numero inicial de elementos.**

   Con base a los resultados obtenidos al hacer pruebas de las dos configuraciones con los factores de carga 2.00 y 8.00 se concluye que para el índice de técnicas es conveniente la configuración Separate Chaining como mecanismo de colisión, con un factor de carga por defecto de 8.00 para manejar un total de 15223 artistas existentes en el archivo MoMa como numero inicial de elementos. Separate Chaining ofrece menores tiempos de carga, lo cual tiene sentido debido al manejo de colisiones que da permitiendo que haya varios sets llave-valor en una misma posición, mientras que en Probing tiene que avanzar a la siguiente posición vacía si la posición original está ocupada, lo cual causa un aumento en el tiempo de ejecución. Sin embargo, al momento de realizar las consultas es más sencillo usar Probing, pues no toca recorrer listas en cada posición que tenga más de un elemento, por lo que terminaremos haciendo uso de linear probing con un factor de carga de aproximado de 0.5 para mantener una buena distribución de los datos y con un número inicial de elementos de 15223.

6. **¿Qué configuración de ADT Map escogería para el índice de nacionalidades?, especifique el mecanismo de colisión, el factor de carga y el numero inicial de elementos.**

   Teniendo en cuenta los resultados hallados en las pruebas de las dos configuraciones con los factores de carga 2.00 y 8.00, es posible afirmar que para la creación del índice de nacionalidades es óptimo utilizar Separate Chaining como mecanismo de colisión y un factor de carga por defecto de 8.00; lo anterior, teniendo en cuenta que se cargarán un total de 15223 artistas del archivo csv. Según las pruebas realizadas para la carga de datos, se espera que con dicha la configuración el tiempo de ejecución tenga un valor cercano a 81156.25ms.

Cabe mencionar que el hecho de que Separate Chaining sea la mejor opción temporal para ambos índices es muy beneficioso debido a que esta configuración además de su buen tiempo de ejecución también ocupara menos espacio en los computadores que se utilicen para correr el programa. Sin embargo, en caso de que se quiera utilizar Lineal Probing para asegurar que no existan datos encadenado en una sola posición y sea más fácil y rápida la consulta de estos, lo mejor sería que este mecanismo fuera acompañado por un factor de carga de 0.5 cuyo tiempo estimado es de 88734.375 ms, pues es un buen intermedio que permite un buen tiempo y, además, permite una mejor distribución de los datos, para que se den menos colisiones y no queden todos concentrados en alguna parte específica del map.