Diego Acosta (202110516)
Mateo Cote (202022609)

Preguntas:

1) Teniendo en cuenta cada uno de los requerimientos ¿Cuántos índices implementaría en el Reto? y ¿Por qué?

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

3) Dado el número de elementos de los archivos MoMA, ¿Cuál sería el factor de carga para estos índices según su mecanismo de colisión?

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?

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 número inicial de elementos.

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

Respuestas:

1. Para este reto se definió implementar aproximadamente 7 índices:
    - Artistas: En este índice se guardaran como llaves los id's de los artistas (cada uno en un espacio de la tabla diferente) y como sus valores respectivos, se guardará el resto de información de este artista
    - Obras: En este índice se guardaran como llaves los id's de cada objeto ("ObjectID") y como valor la información de la obra a la cual pertenece dicho Id
    - Nacionalidad: Este índice será creado para el requerimiento #4 en el cual nos piden catalogar a los artistas por su nacionalidad, para ello se creará este índice donde las llaves serán las nacionalidades y como valores respectivos, una lista de las obras cuyos artistas tengan dicha nacionalidad
    - Medios: Este índice será creado para el requerimiento #3 en donde nos piden catalogar las obras por su técnica. En este caso las llaves serán los nombres de las técnicas y como valores se guardarán listas con las obras que usen dicha técnica
    - Fecha de nacimiento del artista: Este índice será creado para el requerimiento #1, en el cual tendremos que listar los artistas nacidos en cierto rango de años, para esto se creará un índice en el cual las llaves serán los diferentes años de nacimiento y como valores, listas que contienen la información de los artistas nacidos en esos años
    - Fecha de adquisición de la obra: Este índice es necesario para el requerimiento #2 de manera tal que queden organizadas las obras en una tabla por su fecha de adquisición. Por lo tanto, en este índice se cargaran los datos bajo la llave

de la fecha de adquisición al que le corresponde un bucket con todas las obras que fueron adquiridas en dicha fecha.

- Departamento: Este índice es necesario para el requerimiento #5 ya que se deben reconocer cuales son las obras presentes en un departamento para definir su costo de transporte. En este caso, la llave de la tabla será el departamento al que le corresponderá un bucket con todas las obras a transportar.

2. Para definir el tipo de mapa a utilizar para cada uno de los índices se debe identificar la cantidad de datos que se almacena por llave en los diferentes casos. Si a cada espacio de la tabla le corresponde solamente un elemento o valor (si la llave tiene un único valor, por ejemplo un Id hace referencia a un único artista), entonces se utilizará "Linear Probing". Por el contrario, si a cada llave de la tabla le corresponde un bucket de elementos (si cada llave puede tener más de un valor, por ejemplo, una llave de nacionalidad puede tener varias obras cuyos autores tienen esa nacionalidad), se utilizará "Separate Chaining". Bajo este criterio se obtiene que:
    a. Artistas: Linear Probing
    b. Obras: Linear Probing
    c. Nacionalidad: Separate Chaning
    d. Medios: Separate Chaning
    e. Fecha de nacimiento del artista: Separate Chaning
    f. Fecha de adquisición de la obra: Separate Chaning
    g. Departamento: Separate Chaning
3. Hay dos maneras de definir el factor de carga dependiendo del mecanismo de colisión que se vaya a utilizar para el índice. En caso de que se esté utilizando la estructura "Linear Probing", el factor de carga se va a determinar con la fórmula $\frac{n}{m}$donde n es el número de parejas llave-valor y m es el número primo más cercano a el doble de n. De igual manera, para "separate chaining" se utiliza la fórmula $\frac{n}{m}$, pero en este caso la m representa la cantidad de buckets mientras que n se mantendrá igual. Con estos datos preliminares se podrian calcular los factores de carga para los indices:
    a. Artistas: $n = 1949,\ 2n = 3898,\ m = 3907,\ Factor = \frac{1949}{3907} = 0.49$
    b. Obras: $n = 837,\ 2n = 1674,\ m = 1693,\ Factor = \frac{837}{1693} = 0.49$
        i. Para los indices bajo la estructura de "Linear Probing" podemos notar que los valores obtenidos tienden a aproximarse a 0.5 por lo que asumimos que para este mecanismo de colision, se utilizaran dichos factores de carga cercanos a 0.5.
    c. Para los índices bajo la estructura de "Separate Chaining" se presenta un mayor grado de dificultad el tener que identificar la cantidad de buckets ya que se debería agrupar cada uno de los datos por llaves y así determinar cuántos elementos se utilizan para el cálculo del factor de carga. Por lo tanto y teniendo en cuenta la observación sobre el patrón de los factores de carga de linear probing, se puede apostar por factores de carga cercanos a 1.5 (para

poder ver las diferencias con un factor de carga mayor a 1 y notar las diferencias entre mecanismos de colisión).

4. Con los parámetros predeterminados para la estructura “separate chaining” con un factor de carga de 4.0, se obtuvo un promedio de tiempo de ejecución de carga de 57,81 milisegundos (calculado a partir de 5 pruebas). Por otra parte, para la estructura “Linear Probing” con un factor de carga de 0.5, se obtuvo un promedio de tiempo de ejecucion de carga de 61,38 milisegundos. Al cambiar la configuracion se puede observar un aumento en el tiempo de ejecucion cuando se tiene la estructura de “Linear Probing” a diferencia de “Separate chaining”. Es decir, hay una diferencia de aproximadamente 3,57 milisegundos en tiempos de ejecución mostrando así que el mecanismo de colisión más eficiente en este caso, sería “Separate Chaining”.
5. Para el índice de técnicas o medios se escogería la configuración con el mecanismo de colisión “Separate Chaining” con un factor de carga de 2.0 y un número inicial de elementos de 837 (para el archivo “small” de obras) o 150,681 (para el archivo “large” de obras). Esto se debe a que según las pruebas de tiempo de ejecución para la carga de datos, se obtuvo que en orden de eficiencia de mayor a menor: “Separate Chaining 2.0” - 41,97ms, “Separate Chaining 8.0” - 61,28ms, “Linear Probing 0.8” - 62,82ms, “Linear Probing 0.2” - 64,16ms. Esto indica que la configuración que cargará los datos con menor memoria y mayor velocidad será “Separate Chaining 2.0” ya que permite guardar una gran cantidad de obras a una sola llave correspondiente a el medio o técnica.
6. Para el índice de nacionalidades se escogería el mecanismo de colisión de “Separated Chaining” no solo por la facilidad de la organización de los datos (al permitir guardar varias obras o varios ids en una sola llave correspondiente a una nacionalidad), sino porque los tiempos de ejecución son menores a los de un mecanismo “Linear probing”. Los tiempos obtenidos luego de realizar las pruebas de ejecución (en milisegundos) son: “Separate Chaining 2.0” - 187.5, “Separate Chaining 8.0” - 140.625, “Linear Probing 0.8” - 160.875, “Linear Probing 0.2” - 210.75. Por esto, el factor de carga inicial con el que se iniciaría la carga de datos sería de 8.0 con un mecanismo de “Separated Chaining”. El número inicial de elementos 1948 para el archivo small, y 15224 para el large.