Nombres:

Carlos Arturo Holguin Cardenas, 202012385
Daniel Hernández Pineda, 202013995

Preguntas preparatorias

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

En total se tendrían 7 índices.

- Para el primer requerimiento, sería suficiente con crear un índice según "BeginDate" (fechas de nacimiento existentes) cuyos valores serían listas de los artistas nacidos en las respectivas fechas. **1 índice**
- Para el segundo requerimiento, se tendría primero un índice según "ConstituentID" que relacionaría el ID de cada artista con su nombre (esto será útil para más requerimientos). Además, se tendría un índice según "DateAcquired" (fechas de adquisición existentes) cuyos valores serían listas de las obras adquiridas en las respectivas fechas. **2 índices**
- Para el tercer requerimiento, se crearía un índice según "Medium" (técnicas utilizadas en las obras) cuyos valores serían listas de las obras que fueron ejecutadas con las respectivas técnicas. Los artistas se pueden relacionar con sus IDs gracias al índice según "ConstituentID" ya creado. **1 índice**
- Para el cuarto requerimiento, se crearía un índice según "Nationality" (nacionalidades de los artistas) cuyos valores serían listas de las obras relacionadas con las respectivas nacionalidades. Para eso, habría que crear un índice adicional cuyas llaves también fuesen las nacionalidades y sus valores fuesen los IDs de los artistas que pertenecen a las respectivas nacionalidades. **2 índices**
- Para el quinto requerimiento, se crearía un índice según "Department" (departamentos del museo) cuyos valores serían listas de las obras correspondientes al respectivo departamento. **1 índice**

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

Para índices en los que el número de datos sean similares al número de obras, se puede optar por usar ***Separete Chaining,*** ya que puede que se genere un consumo muy alto en memoria. En estos casos, se debe tratar de generar un balance adecuado entre memoria y tiempo, pues gran parte de las máquinas se encuentran limitadas en su memoria RAM. No obstante, si al realizar pruebas de rendimiento se observa que el consumo de memoria es bajo, se puede optar por ***Linear Probing*** para poder disminuir los tiempos.

Lo anterior aplica para el requerimiento 2, puesto que la cantidad de datos de "Constituent ID" es alta. Sin embargo, recordemos que la cantidad de artistas para el archivo "Large" es de 15223 artistas, una cantidad relativamente baja de datos en términos de la capacidad de un computador, por tanto, lo más probable es que se implemente ***Linear Probing***. Otro caso es el del índice según "DateAcquired", el cual,

en el peor caso, puede tener la misma cantidad que número total de obras, pues con fechas tan exactas (AAAA-MM-DD) es probable que existan pocas coincidencias.

En el caso de índices cuyos mapas tendrán tamaños mucho menores que la cantidad total de datos, como lo son “Nationality” (x2), “Department”, “Medium” y “BeginDate”, se daría prioridad al tiempo de ejecución con ***Linear Probing***, ya que la memoria no parecería representar un mayor riesgo para el rendimiento.

c) Dado el número de elementos de los archivos MoMA, ¿Cuál sería el factor de carga para estos índices según su mecanismo de colisión?
Para los archivos “large” hay 138150 obras y 15223 artistas. De esta manera, se partiría de un factor de carga de 0.5 para cuando se use Linear Probing y de un factor de carga de 4.0 cuando se use Separate Chaining. Más adelante, se evaluaría manualmente si es rentable modificar estos factores de carga mediante pruebas de rendimiento.

Separate Chaining o Linear Probing

Separate Chaining con factor de carga de 4

Maquina	**Tiempo (ms)**
PC Carlos Holguin	**6280**
PC Daniel Hernández	**8224**

Linear Probing con factor de carga de 0.5

Maquina	**Tiempo**
PC Carlos Holguin	**6156**
PC Daniel Hernández	**7495**

d) ¿Qué diferencias en el tiempo de ejecución notan al ejecutar la cargar los datos al cambiar la configuración de Linear Probing a Separate Chaining?
En esta primera instancia, el tiempo de ejecución es ligeramente menor al usar Linear Probing.

Modificar el factor de carga

Linear Probing

Maquina	**Tiempo con factor de carga de 0.2 (ms)**	**Tiempo con factor de carga de 0.8 (ms)**
PC Carlos Holguin	**5750**	**6828**
PC Daniel Hernández	**7104**	**8443**

Separate Chaining

Maquina	**Tiempo con factor de carga de 2 (ms)**	**Tiempo con factor de carga de 8 (ms)**
PC Carlos Holguin	**6550**	**7468**
PC Daniel Hernández	**7854**	**8859**

e) ¿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 los datos obtenidos, podemos notar que, al variar el factor de carga, varía considerablemente el tiempo de carga de los datos, por otro lado, se puede observar que en algunos casos se prioriza el tiempo y, por tanto, se genera un mayor consumo en memoria. En este caso lo mejor es optar por **Linear Probing** con un factor de carga de 0.5 y un número inicial de datos de 140.000, esto se debe a que las llaves están asociadas a el **Object ID** de los Artworks. Finalmente, si lo que buscamos es reducir al máximo los tiempos de carga debemos optar por un factor de carga que este alrededor de 0.1 o 0.2, sin embargo, se debe tener presente que se va a generar un mayor consumo de memoria.

f) ¿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.

En este caso, se haría lo mismo, se optaría por usar **Linear Probing** con un factor de carga de 0.5 y un número inicial de datos de alrededor de 500, esto debido a que las llaves están asociadas a la nacionalidad de los autores, por otro lado, sabemos que hay 193 países, por tanto, la tabla de hash está sobredimensionada más de lo necesario, esto se hace para evitar ***rehash.*** Sin embargo, si se quiere disminuir el tiempo se puede optar por un factor de carga que este alrededor de 0.1 o 0.2, sin embargo, se debe tener presente que se va a generar un mayor consumo de memoria.