UTEC
UNIVERSIDAD DE INGENIERÍA
Y TECNOLOGÍA

Welcome to Algorithms and Data Structures! - CS2100

Tablas hash

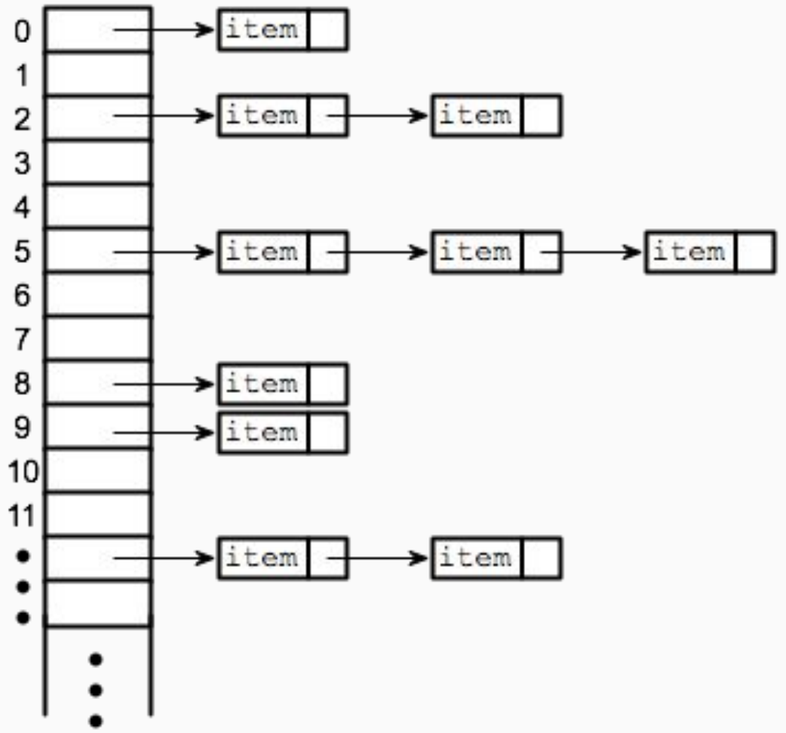
Qué son las tablas hash?
Son estructuras de datos similares a los arrays, con la diferencia que el índice puede ser cualquier tipo de dato comparable y no solo enteros

Restricciones:

- Claves únicas
- No sabemos en qué posición se almacenan los valores

La clave es mapeada a un índice:

```
int index = getIndex(key)
array[index] = value
```

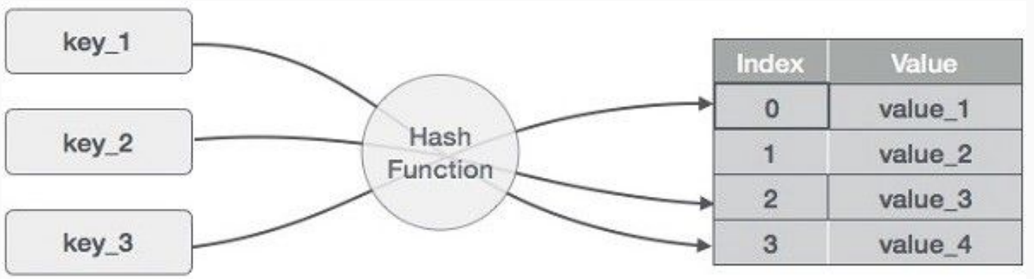


Hashing

Saben lo que es una función hash?
Es un método que nos permite obtener un índice en un array desde una key

Qué problemas podríamos tener con esa función?

- ➔ Procesar la key puede ser difícil para ciertos tipos de datos
- ➔ Manejar colisiones cuando dos keys tienen el mismo índice
- ➔ Espacio-tiempo tradeoff

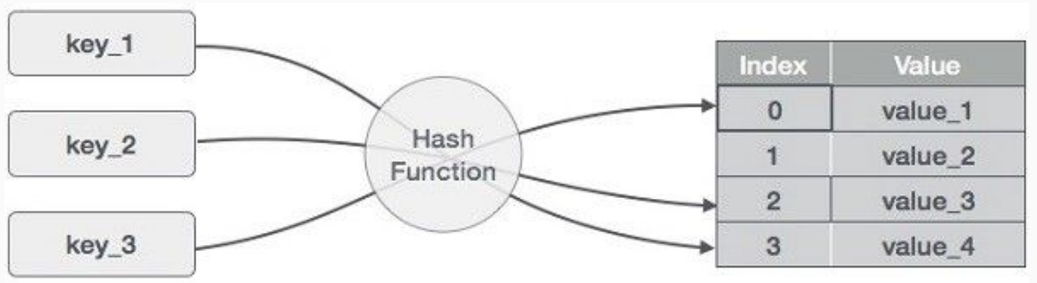


Hashing

Existen varios algoritmos para hashing, algunos retornan un output más grande que otros

Propiedades:

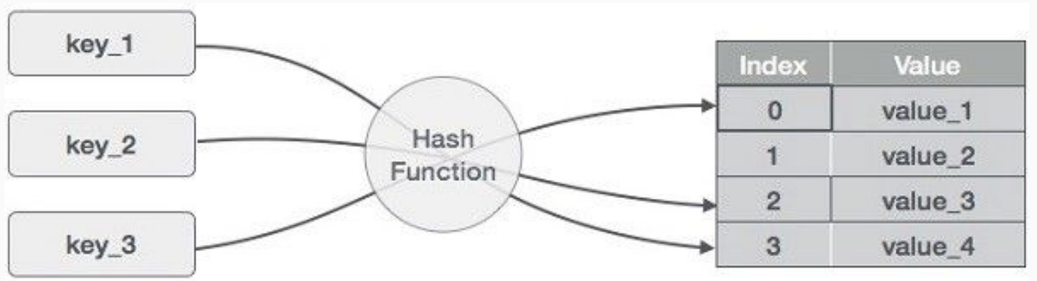
- Estable: El output siempre debería ser el mismo para el mismo input (invariante)
- Uniforme: Los valores hash deben ser distribuidos de manera uniforme (reducir colisiones)
- Eficiencia: Debe ser balanceado de acuerdo a las necesidades en espacio y tiempo
- Seguridad: Dado un valor hash, obtener un valor que me pueda generar ese valor hash no debería ser posible



Hashing (strings)

- Implementación ingenua (suma):

```
void additiveHash(string key) {
   int total = 0;
   for (char& character : key) {
      total += (int) character;
   }
   return total;
}
```

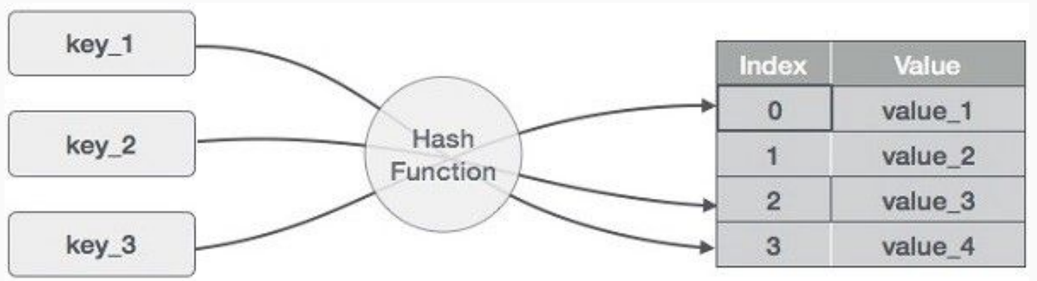


Cuál es el problema de esta función hash?

Hashing (strings)

- Algoritmo plegable:

```
const int getNextBytes(int startIndex, string str) {
  int currentFourBytes = 0;
  currentFourBytes += getByte(str, startIndex);
  currentFourBytes += getByte(str, startIndex + 1) << 8;
  currentFourBytes += getByte(str, startIndex + 2) << 16;
  currentFourBytes += getByte(str, startIndex + 3) << 24;
  return currentFourBytes;
}
```



Cuál es el problema de esta función hash?

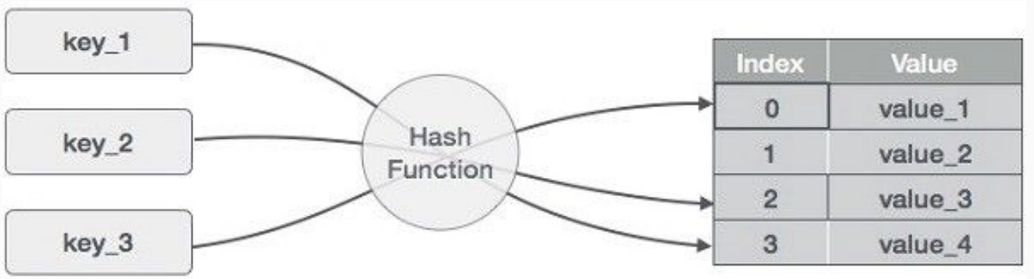
Hashing

No escriban sus propios algoritmos de hash! Utilicen las librerías ya hechas para esto

Escojan el algoritmo hash que encaje mejor con sus necesidades

MD5 solía ser un algoritmo hash bastante usado, tiene un buen balance entre estabilidad y uniformidad pero ya se han detectado colisiones

SHA-2 es un algoritmo bastante seguro pero no es muy eficiente



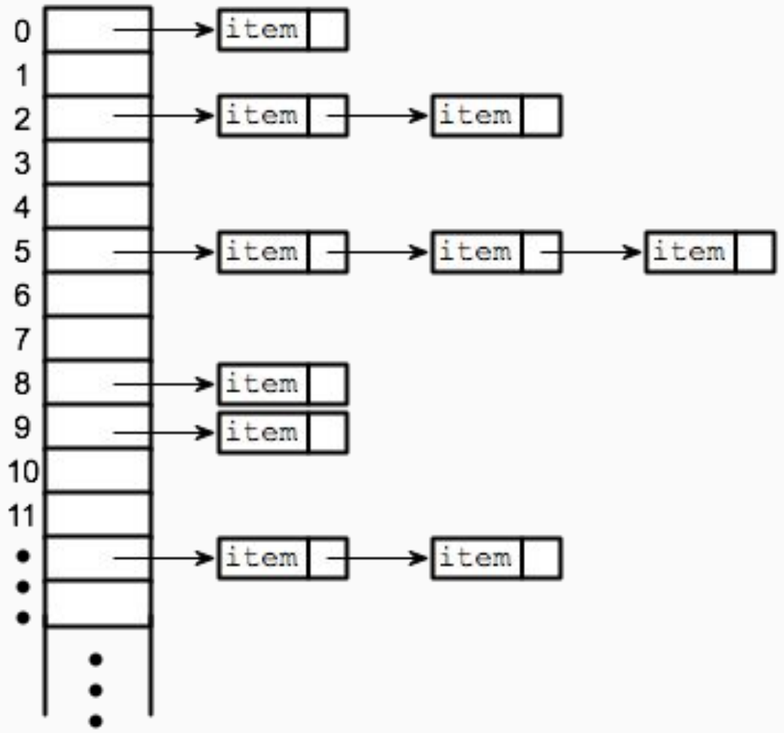
Agregando datos

Se inicia con un tamaño fijo del array, y se genera el hash code

```
int arrayLength = 10;
int hashCode = getHashCode(key);
int index = hashCode % arrayLength;
array[index] = value
```

Qué problemas podrían aparecer?

Dos elementos diferentes con el mismo hash code, significa que se les asigna el mismo índice dentro del array

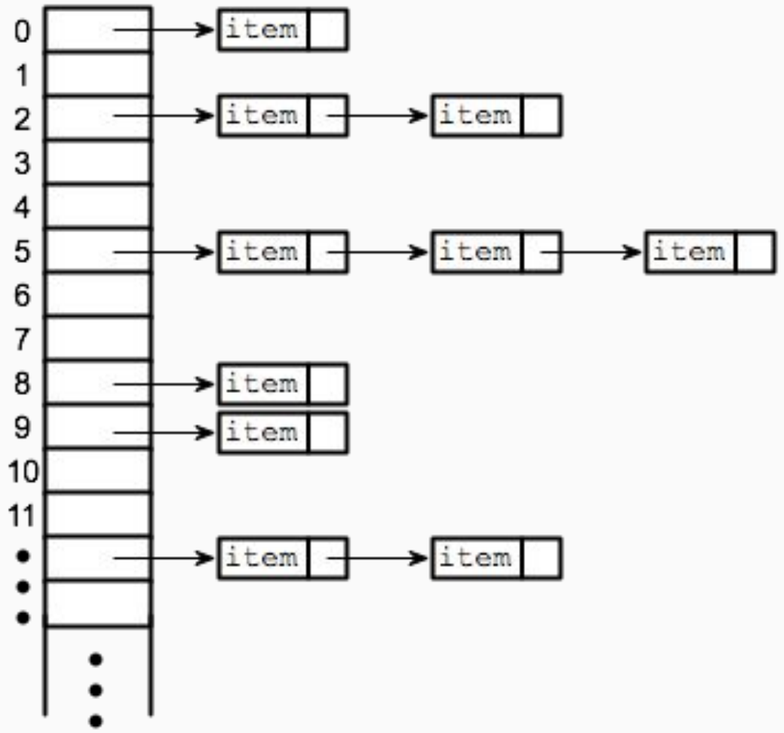


Manejando colisiones

Cómo podrían manejar colisiones?

Recuerden que el usuario de una tabla hash no debe saber que hubo una colisión

- Open addressing: Moverse al siguiente índice disponible
- Chaining: Almacenar los elementos en una lista enlazada

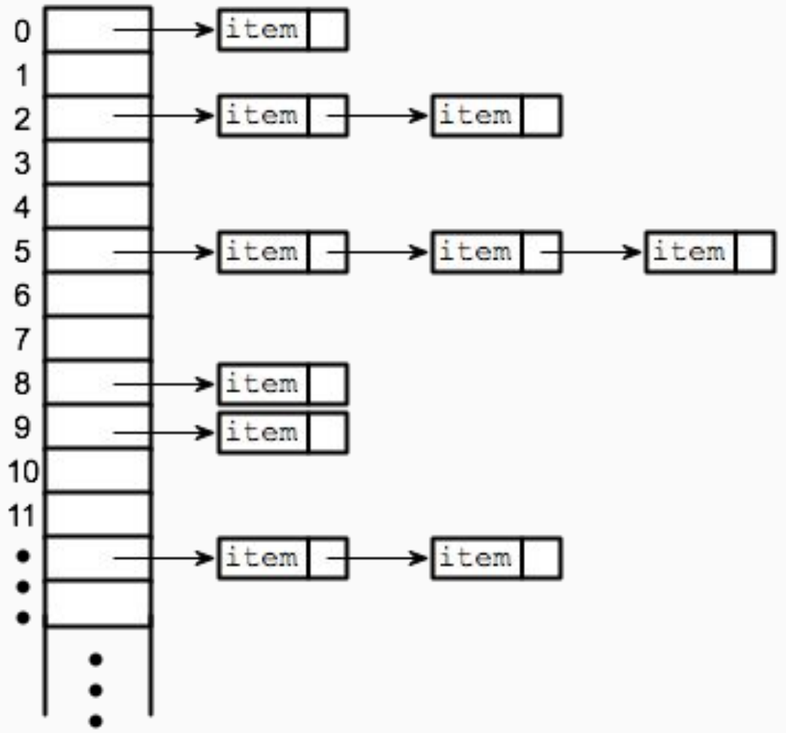


Manejando colisiones (open addressing)

Si el espacio ya está ocupado, entonces se va a buscar el siguiente espacio disponible

```
while (array[index] != null){
    index++;
}
array[index] = value
```

Al momento de buscar el elemento se empieza a comparar el hash code hasta encontrar el elemento o estar seguros que no está en la tabla hash

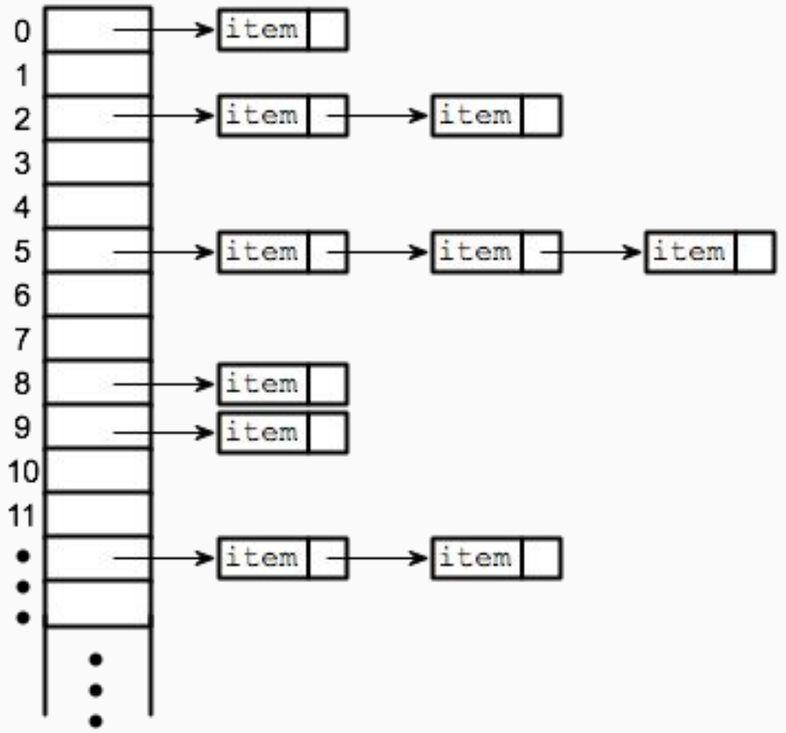


Manejando colisiones (chaining)

Si el espacio ya está ocupado, entonces se agrega el elemento a la lista enlazada

array[index].push_back(value)

Para encontrar el elemento, solo se debe buscar dentro de la lista enlazada



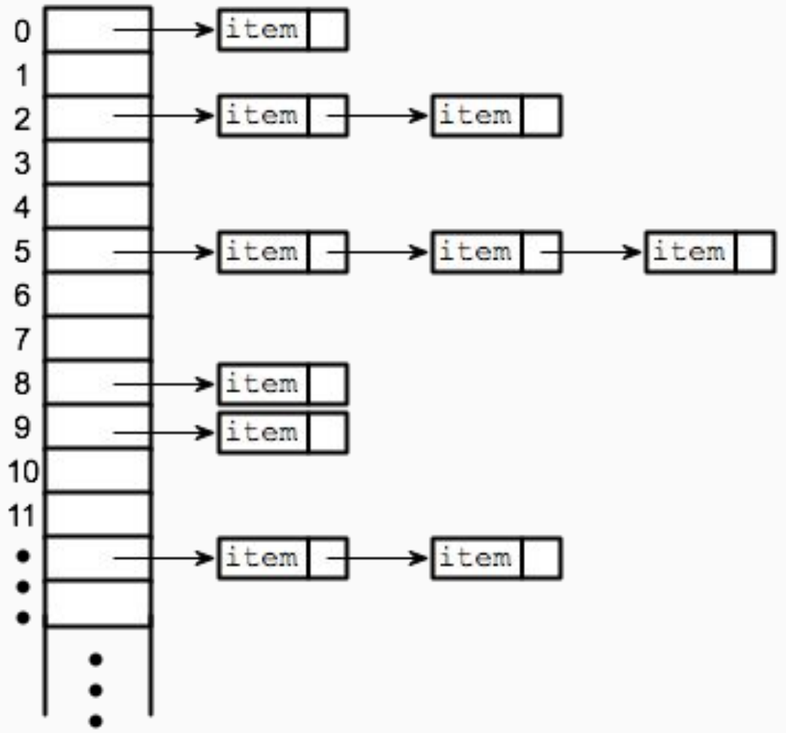
Haciendo crecer la estructura

La probabilidad de que hayan colisiones, va a ser proporcional a la cantidad de espacios disponibles del array

Para controlar esto, se debe tener un factor de llenado, que nos indicará cuán lleno está nuestro hash table.

Al agregar un elemento:

```
if (fillFactor >= maxFillFactor) {
    newArray = new T[2 * array.length()];
    copiarArrayAlNuevoArray();
}
```



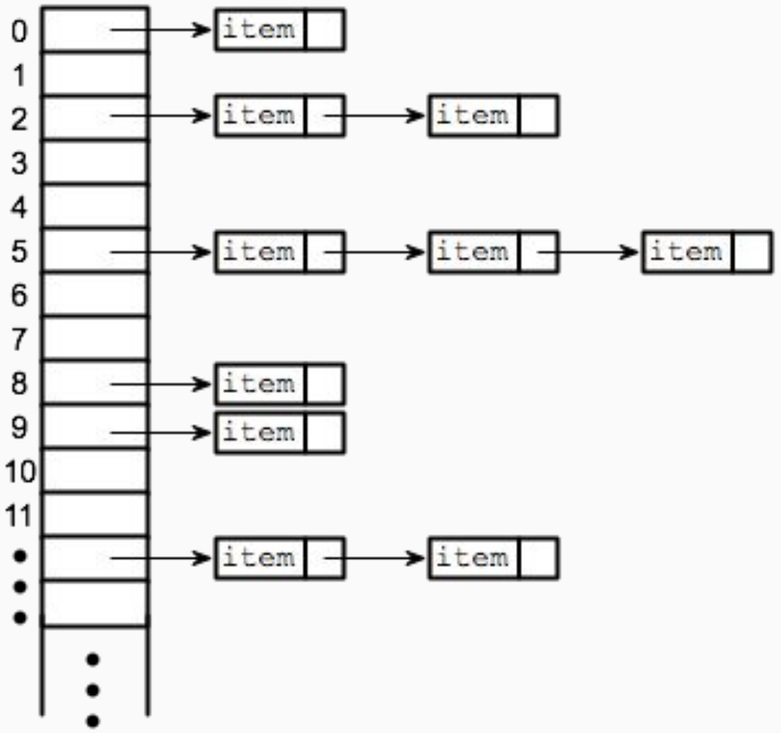
Remover elementos

El remover elementos va a depender de como estemos manejando las colisiones:

Open addressing:

1. Se obtiene el índice del elemento
2. Se verifica que no sea nulo en el array
3. Si las keys coinciden, remover el elemento
4. Si las keys no coinciden (probablemente haya una colisión), revisar el siguiente elemento

Es bastante difícil de mantener

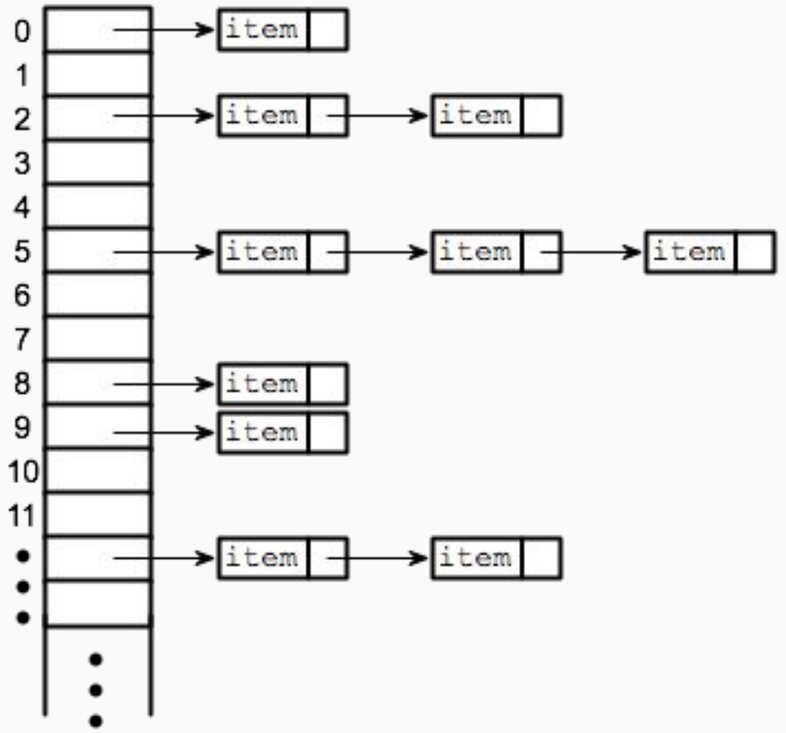


Remover elementos

Chaining:

1. Se obtiene el índice del elemento
2. Se elimina el elemento de la lista simplemente enlazada

Es fácil de mantener e implementar

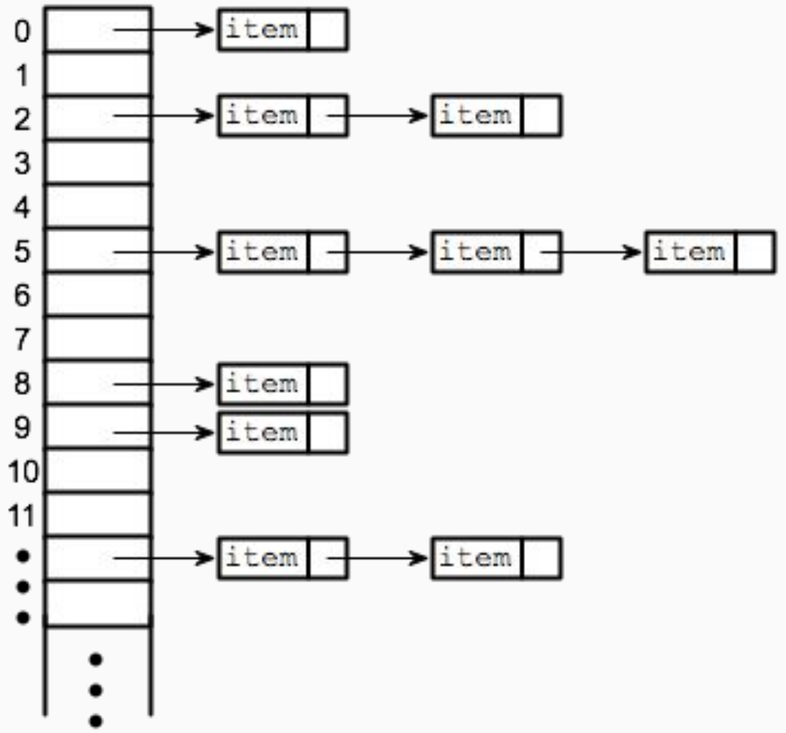


Encontrar elementos

Se aplica la misma lógica que para remover elementos

- Open addressing:
  a. Se obtiene el índice de la key
  b. Se verifica que no sea nulo
  c. Si hay colisión, se sigue avanzando hasta encontrar el elemento

- Chaining:
  a. Se obtiene el índice de la key
  b. Se busca en la lista de la posición



Enumerar elementos

Cuántas maneras de enumerar se deben tener?

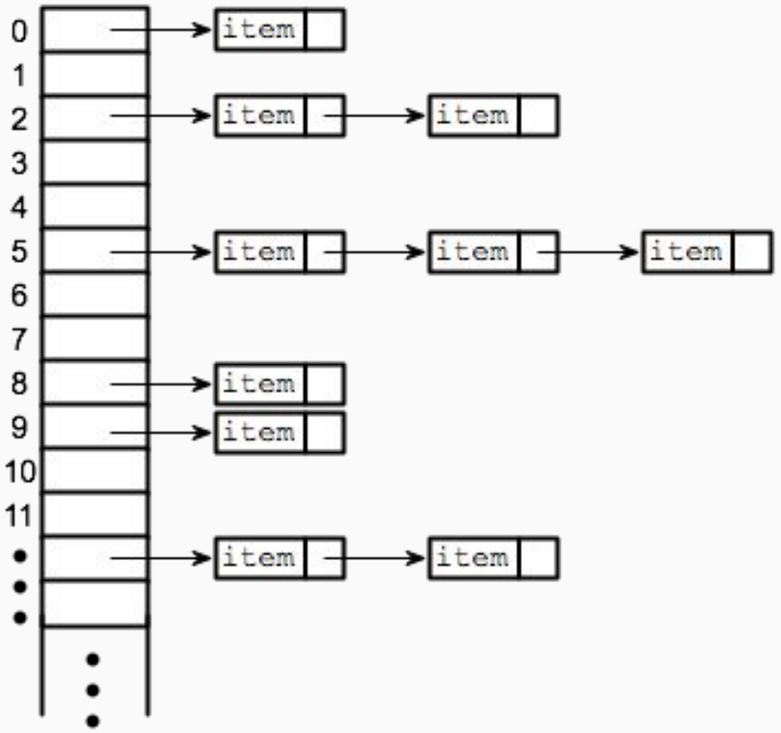
Dos, una para las llaves y otra para los valores

- Open addressing:

```
foreach (item in array) {
    if (item != null) return item;
}
```

- Chaining:

```
foreach(list in array) {
    foreach(item in list) {
        return item;
    }
}
```



UTEC
UNIVERSIDAD DE INGENIERÍA
Y TECNOLOGÍA

Welcome to Algorithms and Data Structures! - CS2100