Info6305: Program Structure & Algorithms

HashTable Report

Abhishek Dongare NUID:001356287

HashTable: Hash table (hash map) is a data structure that implements an associative array abstract data type, a structure that can map keys to values. A hash table uses a hash function to compute an index into an array of buckets or slots, from which the desired value can be found.

Ideally, the hash function will assign each key to a unique bucket, but most hash table designs employ an imperfect hash function, which might cause hash collisions where the hash function generates the same index for more than one key. Such collisions must be accommodated in some way.

In a well-dimensioned hash table, the average cost (number of instructions) for each lookup is independent of the number of elements stored in the table. Many hash table designs also allow arbitrary insertions and deletions of key-value pairs, at (amortized[2]) constant average cost per operation.[3][4]

In many situations, hash tables turn out to be on average more efficient than search trees or any other table lookup structure. For this reason, they are widely used in many kinds of computer software, particularly for associative arrays, database indexing, caches, and sets.

Hashing: The idea of hashing is to distribute the entries (key/value pairs) across an array of buckets. Given a key, the algorithm computes an index that suggests where the entry can be found:

index = f(key, array_size)

hash = hashfunc(key) index = hash % array size Linear Probing:Linear probing is a component of open addressing schemes for using a hash table to solve the dictionary problem. In the dictionary problem, a data structure should maintain a collection of key–value pairs subject to operations that insert or delete pairs from the collection or that search for the value associated with a given key. In open addressing solutions to this problem, the data structure is an array T (the hash table) whose cells T[i] (when nonempty) each store a single key–value pair. A hash function is used to map each key into the cell of T where that key should be stored, typically scrambling the keys so that keys with similar values are not placed near each other in the table. A hash collision occurs when the hash function maps a key into a cell that is already occupied by a different key. Linear probing is a strategy for resolving collisions, by placing the new key into the closest following empty cell.[3][4]

Search

To search for a given key x, the cells of T are examined, beginning with the cell at index h(x) (where h is the hash function) and continuing to the adjacent cells h(x) + 1, h(x) + 2, ..., until finding either an empty cell or a cell whose stored key is x. If a cell containing the key is found, the search returns the value from that cell. Otherwise, if an empty cell is found, the key cannot be in the table, because it would have been placed in that cell in preference to any later cell that has not yet been searched. In this case, the search returns as its result that the key is not present in the dictionary.[3][4]

Insertion

To insert a key–value pair (x,v) into the table (possibly replacing any existing pair with the same key), the insertion algorithm follows the same sequence of cells that would be followed for a search, until finding either an empty cell or a cell whose stored key is x. The new key–value pair is then placed into that cell.[3][4]

If the insertion would cause the load factor of the table (its fraction of occupied cells) to grow above some preset threshold, the whole table may be replaced by a new table, larger by a constant factor, with a new hash function, as in a dynamic array. Setting this threshold close to zero and using a high growth rate for the table size leads to faster hash table operations but greater memory usage than threshold values close to one and low growth rates. A common choice would be to double the table size when the load factor would exceed 1/2, causing the load factor to stay between 1/4 and 1/2.

Double Hashing: Double hashing with open addressing is a classical data structure on a table {\displaystyle T} T. Let {\displaystyle n} n be the number of elements stored in {\displaystyle T} T, then {\displaystyle T} T's load factor is {\displaystyle \alpha = {\frac {n}{|T|}}} \alpha = \frac{n}{|T|}.

Double hashing approximates uniform open address hashing. That is, start by randomly, uniformly and independently selecting two universal hash functions ${\displaystyle h_{1}} h_{1} \ dsplaystyle \ h_{2} \ dsplaystyle \ T$.

All elements are put in ${\displaystyle T} T$ by double hashing using ${\displaystyle h_{1}} h_{1}$ and ${\displaystyle h_{2}} h_{2}$. Given a key ${\displaystyle k} k$, determining the ${\displaystyle (i+1)} (i+1)$ -st hash location is computed by:

 ${\displaystyle \{(i,k)=(h_{1}(k)+i\cdot h_{2}(k))\cdot h(i,k)=(h_{1}(k)+i\cdot h_{2}($

Let {\displaystyle T} T have fixed load factor {\displaystyle \alpha :1>\alpha >0} \alpha: 1 > \alpha > 0. Bradford and Katehakis[1] showed the expected number of probes for an unsuccessful search in {\displaystyle T} T, still using these initially chosen hash functions, is {\displaystyle {\frac {1}{1-\alpha }}} \frac{1}{1-\alpha} regardless of the distribution of the inputs. More precisely, these two uniformly, randomly and independently chosen hash functions are chosen from a set of universal hash functions where pairwise independence suffices.

Previous results include: Guibas and Szemerédi[2] showed {\displaystyle {\frac {1}{1-\alpha }}} \frac{1}{1-\alpha} holds for unsuccessful search for load factors {\displaystyle \alpha <0.319} \alpha < 0.319. Also, Lueker and Molodowitch[3] showed this held assuming ideal randomized functions. Schmidt and Siegel[4] showed this with {\displaystyle k} k-wise independent and uniform functions (for {\displaystyle k=c\log n} k = c \log n, and suitable constant {\displaystyle c} c).

Load Factor:

A critical statistic for a hash table is the load factor, defined as

LoadFactor = N / K.

where

n is the number of entries occupied in the hash table. k is the number of buckets. As the load factor grows larger, the hash table becomes slower, and it may even fail to work (depending on the method used). The expected constant time property of a hash table assumes that the load factor is kept below some bound. For a fixed number of buckets, the time for a lookup grows with the number of entries and therefore the desired constant time is not achieved. As a real-world example, the default load factor for a HashMap in Java 10 is 0.75, which "offers a good tradeoff between time and space costs." [10]

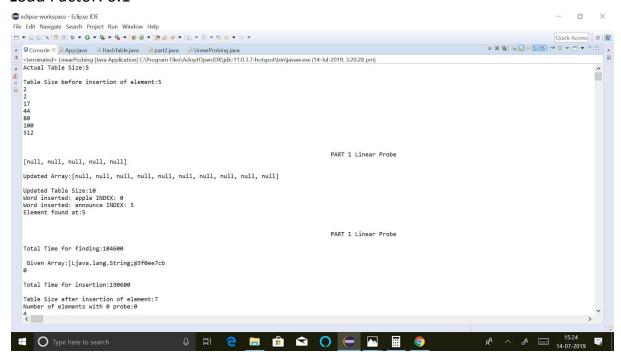
Second to that, one can examine the variance of number of entries per bucket. For example, two tables both have 1,000 entries and 1,000 buckets; one has exactly one entry in each bucket, the other has all entries in the same bucket. Clearly the hashing is not working in the second one.

A low load factor is not especially beneficial. As the load factor approaches 0, the proportion of unused areas in the hash table increases, but there is not necessarily any reduction in search cost. This results in wasted memory.

Empirical Analysis:

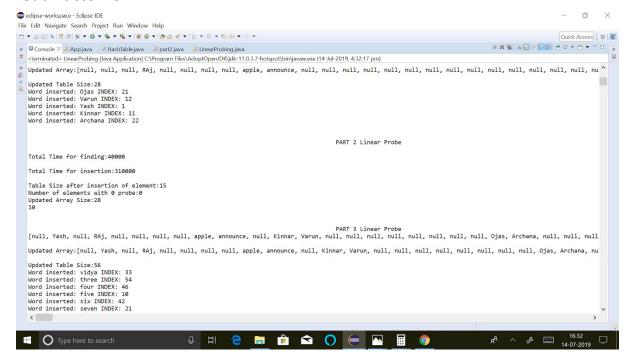
Linear probing:

Load Factor: 0.1



Time for insertion: 190600 Nano seconds. Time for finding: 104600 Nano seconds.

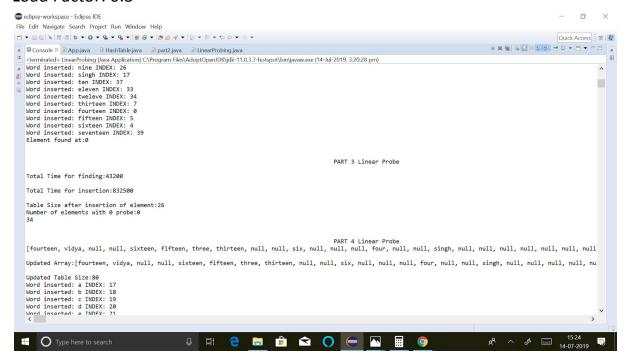
Load Factor: 0.2



Time for insertion:310000 Nano seconds.

Time for finding: 40000 Nano seconds.

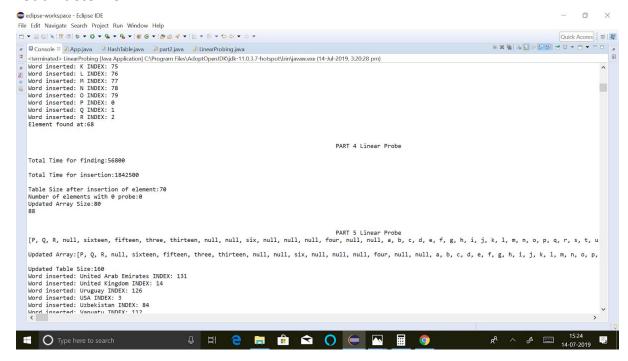
Load Factor: 0.3



Time for insertion: 832500Nano seconds.

Time for finding: 43200 Nano seconds.

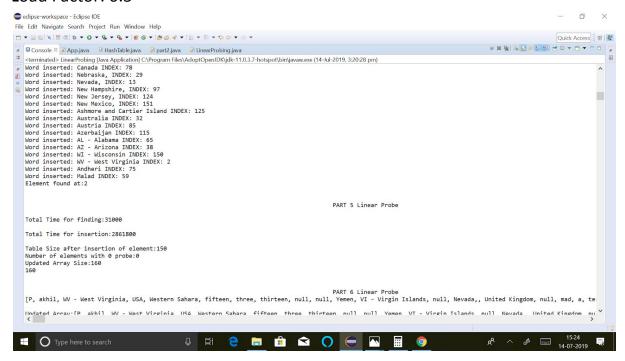
Load Factor: 0.4



Time for insertion: 1842500 Nano seconds.

Time for finding: 56800 Nano seconds.

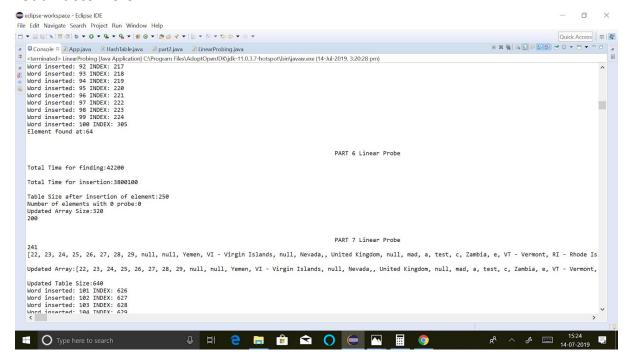
Load Factor: 0.5



Time for insertion: 2861800 Nano seconds.

Time for finding: 31000 Nano seconds.

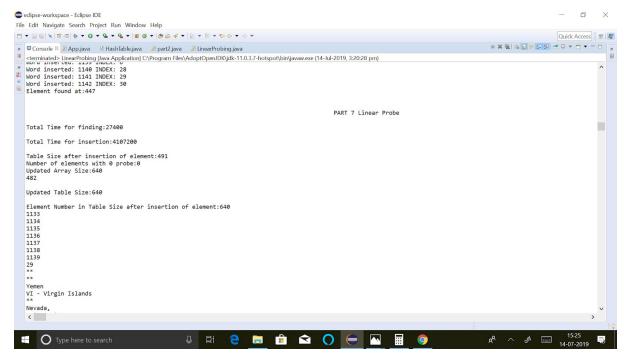
Load Factor: 0.6



Time for insertion: 3800100 Nano seconds.

Time for finding: 42200 Nano seconds.

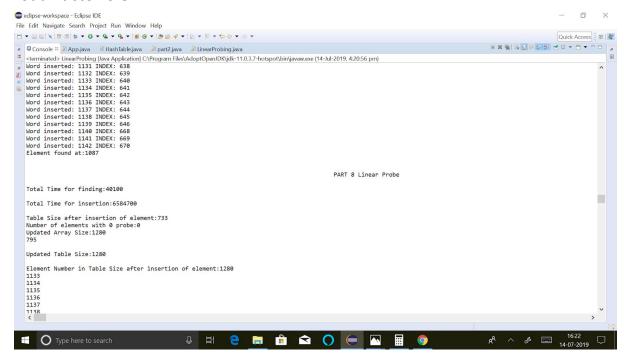
Load Factor: 0.75



Time for insertion: 4107200 Nano seconds.

Time for finding: 27400 Nano seconds.

Load Factor:0.8



Time for insertion: 6584700 Nano seconds.

Time for finding: 40100 Nano seconds.

Load Factor	Insertion Time	Finding time
0.1	190600	104600
0.2	310000	40000
0.3	832500	43200
0.4	1842500	56800
0.5	2861800	31000
0.6	3800100	42200
0.75	4107200	27400
0.8	6584700	40100

Double Hash:

Double hashing with open addressing is a classical data structure on a table ${\displaystyle T} T$. Let ${\displaystyle n} n$ be the number of elements stored in ${\displaystyle T} T$, then ${\displaystyle T} T$'s load factor is ${\displaystyle \alpha = {frac{n}{|T|}} \alpha = {frac{n}{|T|}}.$

Double hashing approximates uniform open address hashing. That is, start by randomly, uniformly and independently selecting two universal hash functions ${\displaystyle \frac{1}{h_{1}} h_{1} }$ and ${\displaystyle \frac{2}{h_{2}} h_{2} }$ to build a double hashing table ${\displaystyle \frac{1}{h_{1}} }$ T.

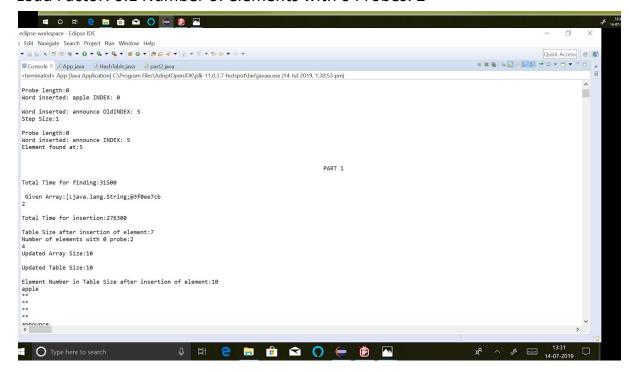
All elements are put in ${\displaystyle \frac{h_{1}} h_{1} \ displaystyle T} T \ by double hashing using <math>{\displaystyle \frac{h_{1}} h_{1} \ displaystyle h_{2}} h_{2}$. Given a key ${\displaystyle \frac{k} k, \ determining the {\displaystyle \frac{(i+1)} (i+1)-st \ hash \ location is computed by:} }$

 ${\displaystyle (h_{1}(k)+i \cdot h_{2}(k)) \pmod |T|.} \ h(i,k) = (h_{1}(k)+i \cdot h_{2}(k)) \pmod |T|.}$

Let {\displaystyle T} T have fixed load factor {\displaystyle \alpha :1>\alpha >0} \alpha: 1 > \alpha > 0. Bradford and Katehakis[1] showed the expected number of probes for an unsuccessful search in {\displaystyle T} T, still using these initially chosen hash functions, is {\displaystyle {\frac {1}{1-\alpha }}} \frac{1}{1-\alpha} regardless of the distribution of the inputs. More precisely, these two uniformly, randomly and independently chosen hash functions are chosen from a set of universal hash functions where pair-wise independence suffices.

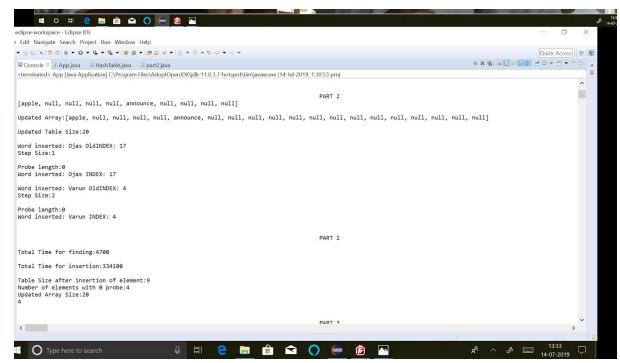
Previous results include: Guibas and Szemerédi[2] showed {\displaystyle {\frac {1}{1-\alpha }}} \frac{1}{1-\alpha} holds for unsuccessful search for load factors {\displaystyle \alpha < 0.319} \alpha < 0.319. Also, Lueker and Molodowitch[3] showed this held assuming ideal randomized functions. Schmidt and Siegel[4] showed this with {\displaystyle k} k-wise independent and uniform functions (for {\displaystyle k=c\log n} k = c \log n, and suitable constant {\displaystyle c} c).

Load Factor: 0.1 Number of elements with 0 Probes: 2



Time for insertion: 276300 Nano seconds. Time for finding: 31500 Nano seconds.

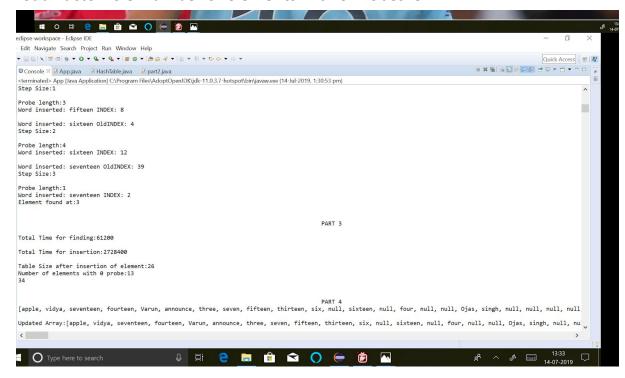
Load Factor: 0.2 Number of elements with 0 Probes: 2



Time for insertion: 334100 Nano seconds.

Time for finding: 4700 Nano seconds.

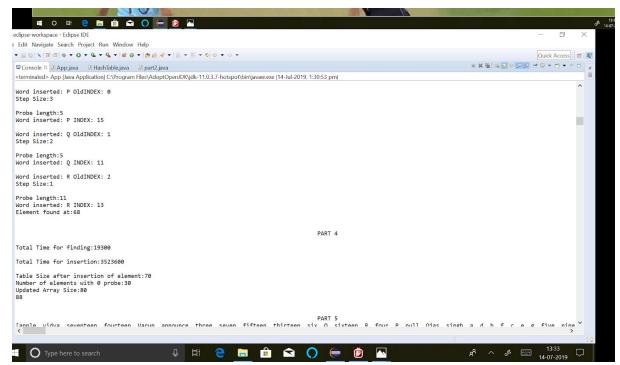
Load Factor: 0.3 Number of elements with 0 Probes: 9



Time for insertion: 2728400 Nano seconds.

Time for finding: 61200 Nano seconds.

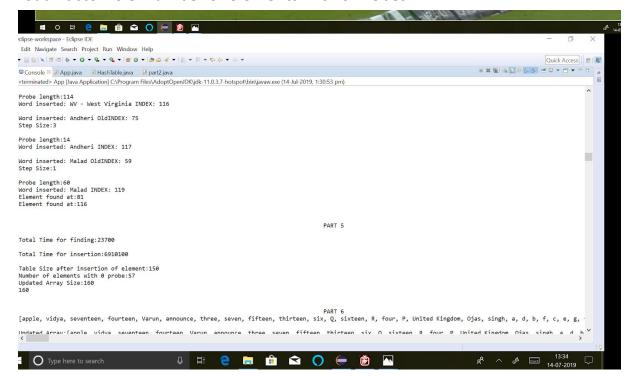
Load Factor: 0.4 Number of elements with 0 Probes: 17



Time for insertion: 3523600 Nano seconds.

Time for finding: 19300 Nano seconds.

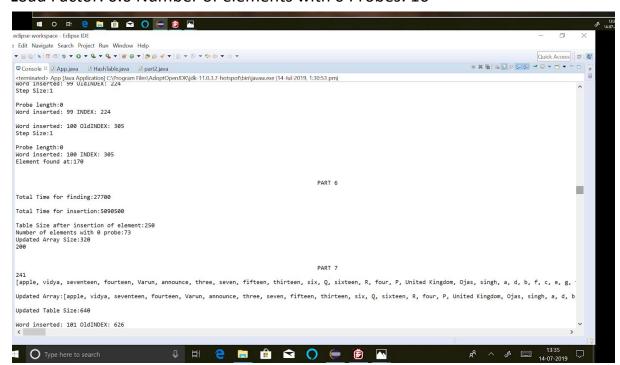
Load Factor: 0.5 Number of elements with 0 Probes: 27



Time for insertion: 6910100 Nano seconds.

Time for finding: 23700 Nano seconds.

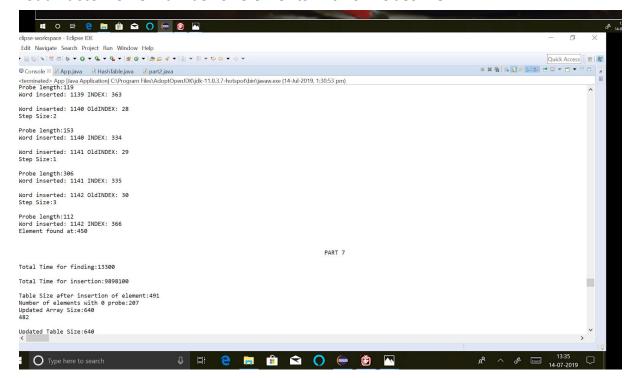
Load Factor: 0.6 Number of elements with 0 Probes: 16



Time for insertion: 5090500 Nano seconds.

Time for finding: 27700 Nano seconds.

Load Factor: 0.75 Number of elements with 0 Probes: 134

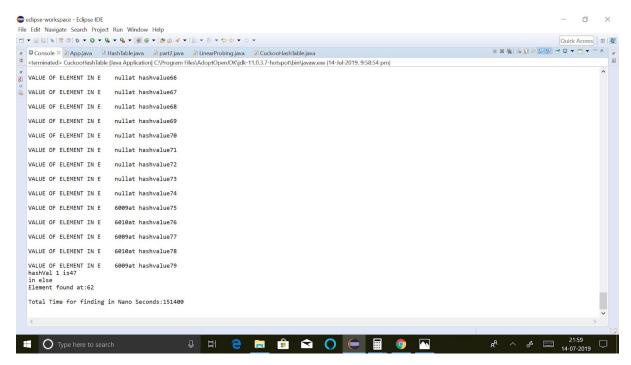


Time for insertion: 9898100 Nano seconds.

Time for finding: 13300 Nano seconds.

Load Factor	Insertion Time	Finding time
0.1	276300	31500
0.2	334100	4700
0.3	2728400	61200
0.4	3523600	19300
0.5	6910100	23700
0.6	5050500	27700
0.75	9898100	13300
0.8	6584700	40100

CUCKOO HASHING:



It can be seen from the above observation that time required to find an element using Cuckoo Hash is more. Hence Cuckoo hash takes lot of time in linear probing.

Conclusion:

Number of elements with 0 probes indicate the number of elements which were added to their respective indices after resizing. This number specifies the efficiency of the hashtable after resizing. Thus at load factor 0.75, we find the element with the least time and also the number of elements with 0 probe is maximum at 0.75. Thus for the above empirical analysis, 0.75 is the breakeven point of the algorithm.