11 Feb 2026

Introducing Hashing

Announcement. NO QUIZ on Wed., Feb. 18 due to Feb. break.

Dictionary Hash Table Hash Function

Hash Map

Dictionary. An abstract data type that stores values assoc. with keys. Operations are:

- INSERT(k,v): inserts value v with key k
- LOOKUP(k): retrieves value stored with k
- DELETE(k): removes k and its value from the dictionary.

Hash Table == Hash Map: a particular implementation of dictionary.

Hash Function. A random function used internally to the hash table.

buckets, hash buckets.

Hash function. A random function $h: \mathcal{X} \to \mathcal{B}$ that yields

keys/ inputs

consistent answers when queried on the same input.

Ideally we would like 3 properties.

(a priori before h sampled)

(1.) Uniform randomness. $\forall k \in \mathbb{N}$

For any set of distinct inputs $(x_1, \ldots, x_k)$ the random k-tuple $(h(x_1), h(x_2), \ldots, h(x_k))$ is uniformly distributed over $\mathcal{B}^k$.

(2) Reproducibility. (a posteriori after h sampled)

Every time you query $h(x)$ on the same input, x, it evaluates to the same answer.

(3) Space/time efficiency.

Ideally evaluating $h(x)$ takes $O(1)$ time and storing the description of h takes $O(\log |\mathcal{B}|)$ space.

<u>CONVENTION.</u> There is a "problem scale parameter", n. (Input size we anticipate.)

$$|\mathcal{B}| \leq n^{O(1)}$$

$$|X| \leq n^{O(1)}$$

One unit of storage is $O(\log n)$ bits.

Arithmetic ops, dereferencing pointers, comparing $x \stackrel{?}{==} y$ take $O(1)$ time when args are $O(\lg n)$ bits.

Fact. Any hash function impl that satisfies

- uniform randomness
- reproducible answers on up to m distinct inputs

must use $\Omega(m)$ space.
(i.e. $\Omega(m \log_2 n)$ bits of storage)

Proof. Consider m distinct inputs $x_1, \dots, x_m$. Say $|B| = n$.
Sequence of operations:

$$h(x_1)\ h(x_2) \dots h(x_m) \mid h(x_1)\ h(x_2) \dots h(x_m).$$

Checkpoint

Claim. The data structure has at least n^m internal states it could be in when checkpt. is reached.

For each $\vec{b} \in B^m$
$= (b_1, \dots, b_m)$

let $S(b) = \{$ potential internal states on executions where $h(x_1) = b_1, \dots, h(x_m) = b_m \}$.

① $S(b) \neq \emptyset \quad \forall\, b \in B^m$.
(b/c uniform randomness)

② If $b \neq b'$ $\quad S(b) \cap S(b') = \emptyset$.

then let $s \in S(b) \cap S(b')$.
$\exists$ two computation paths leading to state s with distinct answer sequences before checkpt.

The answers to the m queries following the checkpt. cannot match <u>both</u> b and b'

$\therefore$ $\Pr(\text{violating reproducibility}) > 0$.

Conclusion:

(unif rand) + (reprod.)

$\implies$ {potential internal states}

has n^m disjoint non-empty subsets.

$|\text{States}| \geq n^m$

bits of storage $\geq m \log_2 n$

Space Complexity $\geq m$.

Def. Hash function distribution $\mathcal{D}$ is k-universal if $\forall$ distinct $(x_1, \ldots, x_k)$ the distrib. of $h(x_1), \ldots, h(x_k)$ is uniform over $\mathcal{B}^k$.

Ex. Say $|\mathcal{B}| = p$ prime, and say $|\mathcal{X}| = p$.

Identify $X = B = \mathbb{Z}/(p)$ (integers mod p).

Sampling h: pick $a, b \in \mathbb{Z}/(p)$ uniformly at random.

Store a, b. (Description of h.)

$$h(x) = ax + b \pmod{p}.$$

"Linear congruential"

If $x \neq y$ there are p^2

ways to sample a, b.

There are p^2 values for

$(h(x), h(y))$.

Claim: The function

$$(a, b) \mapsto (ax+b, ay+b) \bmod p$$

is injective.

Suppose

$$\begin{aligned} ax + b &= a'x + b' \\ ay + b &= a'y + b' \end{aligned} \pmod{p}$$

$$a(x-y) = a'(x-y) \pmod{p}$$

$$(a-a')\cdot\cancel{(x-y)}\equiv 0 \pmod p$$

$$\neq 0$$

$$a\equiv a' \mod p.$$

$$a\cancel{x}+b\equiv a'\cancel{x}+b' \pmod p$$